Efficient estimation of choice-based sample methods with the method of moments

Imbens, G.W.

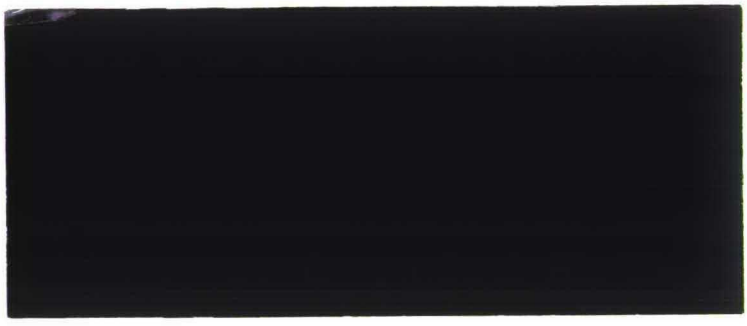
Publication date:
1989

Document Version
Publisher's PDF, also known as Version of record

ERSITY

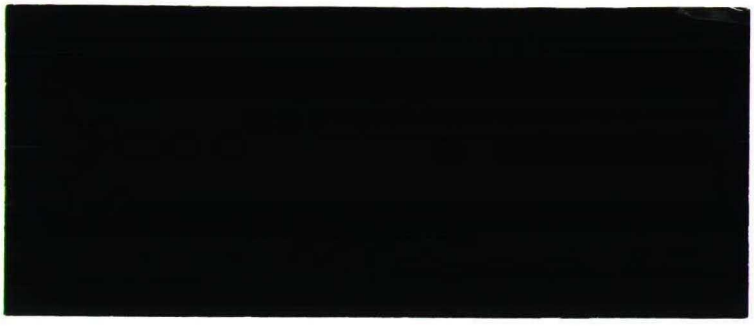
VERSITEIT
BRABANT

POSTBOX 90153
5000 LE TILBURG
THE NETHERLANDS



DEPARTMENT OF ECONOMICS

RESEARCH MEMORANDUM



EFFICIENT ESTIMATION OF CHOICE-BASED SAMPLE MODELS WITH THE METHOD OF MOMENTS

Guido W. Imbens

FEW 417

Efficient Estimation of Choice–based Sample Models with the Method of Moments

Guido W. Imbens[1]
Brown University and Tilburg University[2]

November 16, 1989

[1]I wish to acknowledge stimulating discussions with Tony Lancaster and helpful remarks from Bertrand Melenberg and the participants in a seminar at University College London

[2]Mailing address: Tilburg University, Postbus 90153, 5000 LE Tilburg, The Netherlands

Abstract

In this paper a new estimator is proposed for discrete choice models with choice–based sampling. Existing estimators suffer from a number of disadvantages. Several estimators are not efficient while those that are, are notoriously hard to compute. Another unappealing characteristic of some of these estimators, including those that are efficient, is that if one replaces some of the parameters of the optimatization program by their probability limits, one estimates the resulting parameters with less accuracy. The new estimator is efficient while relatively easy to compute. Its form also sheds light on the causes of the aforementioned counter–intuitive results.

1 Introduction

In this paper a unified theory will be presented for estimating parameters of discrete choice models with choice–based samples. Discrete choice models, or qualitative response models as they are also called, are characterized by the feature that the dependent variable is discrete instead of continuous. Examples are modes of transport, choices of school types or participation decisions.

Sometimes a number of the alternatives are very rare while still important to the researcher. Incidence of rare diseases, or the choice of a particular school type are examples. In that case the researcher might want to oversample that particular response to increase the accuracy of his analysis (be it the estimation of parameters or the prediction of behaviour). Especially in dynamic models it often happens that responses, in this case life histories, that contain relatively much information, occur relatively infrequently. See for a discussion of choice–based sampling in a dynamic context Ridder [27] and Lancaster and Imbens [19]. Another area where this is relevant is that of evaluation of training schemes, discussed in, among others, Heckman and Hotz [16]. If the conventional practice of specifying the conditional distribution of the dependent variable rather than the joint distribution of the dependent and the independent or explanatory variables is maintained, standard maximum likelihood techniques do not apply. It is this case that is the subject of the choice–based, response–based or endogenous sampling literature.

In this paper an estimator is proposed that improves on those that have been suggested previously. Some of these earlier estimators such as those by Manski and Lerman, and Manski and McFadden are inefficient, while the ones that are efficient, notably those proposed by Cosslett are very hard to compute. The new estimator has the same efficiency as those by Cosslett but reduces the computational burden. All three of the aforementioned estimators have a common unappealing and counter intuitive feature. They are defined as solutions to equations which contain additional parameters. Substituting the true values or probability limits of these nuisance parameters into those equations, rather than estimating them jointly with the parameters of interest reduces the efficiency of the estimator for the parameters of interest. The form of the new estimator sheds some light on

the nature of this anomaly.

The estimators that have been suggested in the literature can be divided into two groups, firstly those that assume that the populations probabilities of the choices are known and secondly those that assume that they are not. The new estimator incorporates these two extremes as special cases and can cope with partial knowledge of the probabilities. If these probabilities are known they give rise to stochastic restrictions on the other parameters that can be treated as moment equations. If they are not known, they will be treated as additional parameters and estimated using the same equations that are used as stochastic restrictions in the other case.

The procedure followed to obtain the estimator and the form that is eventually derived, provide some intuition about the way in which information about the marginal distribution of the dependent variable can be used efficiently. It is similar to the procedure used by Chamberlain [5,6] to prove efficiency of method of moments estimators. First it is assumed that the exogenous variables have a discrete distribution with known points of support. In that case one can estimate the parameters of interest by Maximum Likelihood techniques. The next, crucial step is to change the estimator thus obtained into one that is valid whatever the distribution of the exogenous variables. The functions that can be interpreted as score functions in the Maximum Likelihood framework will be interpreted as moment functions in the Method of Moments framework.

The result is a simpler estimator for the case where the population proportions are known in the sense that optimization takes place over a space of lower dimension. This is important because the computational difficulties with Cosslett's estimators are severe as noted by Cosslett [9], Manski and McFadden [23] and Gourieroux and Montfort [12]. Specification tests based on the population proportions are also provided.

The plan of the paper is as follows: in section 2 the issues in choice–based sampling are formally stated and the solutions from the literature are discussed. In section 3 the new estimator is developed and its properties analyzed. A summary and conclusion are given in section 4.

2 Notation and Previous Estimators

In the first subsection the notation will be set up. This is a complicated matter in the choice-based sampling literature. For every random variable one has not only the population distribution and its sample equivalent, the empirical distribution or sample frequency, but also the distribution according to which the data are drawn.

The first one, the population distribution, is what one is interested in. The second one, the sample frequency is known and has to be used to learn about the first. The last one will sometimes be labelled *sample distribution* in this paper, a term that indicates that it is somewhere between the *population distribution* and the *sample frequency*. From the data one can learn the sample frequency and eventually about the sample distribution. Identification refers to the possibility to infer the parameters of the population distribution from (those of) the sample distribution.

If the sample were random, the sample distribution would be identical to the population distribution and it need not be distinguished from it. If the sampling were *exogenous*, i.e. the sampling depends on the values of the exogenous variables, then the sampling distribution does differ from the population distribution but it does not matter. It is the fact that it does matter in the endogenous sampling case that makes the notation more difficult.

In the subsequent subsections three estimators that have been proposed in the literature are discussed. The first is the *weighted exogenous sampling maximum likelihood* estimator. Its form is not of particular relevance for the new estimator proposed later but the generality of the approach behind the WESML estimator and some new results on its relative efficiency warrant its inclusion here. The second estimator discussed is the *conditional maximum likelihood* estimator. It is important for the discussion as we will be able to locate the source of inefficiency for this estimator very clearly. A slightly different form of the CML estimator will have scores that are identical to some of the moments of the new estimator. The last estimators discussed in this section are due to Cosslett. These are the estimators that the paper tries to improve upon in terms of computational ease and intuition.

2.1 Notation

In a population the joint density of a discrete random variable i and a continuous[1], vector valued random variable x is

$$(1) \quad f(i,x) = P(i|x,\theta)r(x)$$

for $i \in C = \{1,2,\ldots,M\}$, $x \in \mathcal{X} \subset \Re^L$ and $\theta \in \Theta \subset \Re^K$. The distribution function of x will be denoted by $R(x)$. We are interested in the parameter θ of the conditional probabilities. One might also be interested in $Q(i)$, the marginal probability or population share of choice i. Even if one is not interested in $Q(i)$ itself, it is useful to define it explicitly. This will make it easier to incorporate prior information about it and such prior information (namely that one of the choices is very rare) was one of the motivations for sampling choice-based. In fact, early studies on choice-based sampling as Manski and Lerman [22] focused exclusively on the case where these probabilities are known exactly. The true value of θ is θ^* and the corresponding notation for $Q(i)$ is $Q^*(i)$:

$$(2) \quad Q^*(i) = \int_{\mathcal{X}} P(i|x,\theta^*)dR(x)$$

Observations are not drawn randomly from this population. With probability H_s an observation is drawn randomly from that part of the population for which $i \in \mathcal{J}(s) \subset C$, $\mathcal{J}(s) \neq \emptyset$ for all $s = 1,2,\ldots S$. The H_s satisfy $\sum_{s=1}^{S} H_s = 1$, $H_s > 0$. At times these probabilities of sampling from the different subpopulations or strata will be assumed not to be known to the investigator. In that case H_s^* will denote true values. The $S-1$ dimensional vector $(H_1\ H_2\ \ldots\ H_{S-1})$ will be denoted by H and the $M-1$ dimensional vector $(Q(1)\ Q(2)\ \ldots\ Q(M-1))$ by Q. H_S and $Q(M)$ will be used as shorthand for $1 - \sum_{t=1}^{S-1} H_t$ and $1 - \sum_{j=1}^{M-1} Q(j)$ respectively.

Each choice i can be in zero, one or more of the subpopulations. The number of strata of which it is a member is denoted by S_i:

$$(3) \quad S_i = \sum_{s=1}^{S} I[i \in \mathcal{J}(s)]$$

[1] the continuity assumption is not essential

where $I[.]$ is the indicator function, equal to one if the expression between the brackets is true and zero otherwise. Examples of sampling strategies included in the class defined above implicitly are:

1. $S = 1$, $\mathcal{J}(1) = C$. This is just random sampling. Standard Maximum Likelihood techniques apply.

2. $S = 2$, $\mathcal{J}(1) = C$, $\mathcal{J}(2) = \{1\}$ In this case the first subsample is completely random, or, in other words, the first stratum is equal to the population. The second subsample consists of observations with choice 1. This is often called an *augmented* sample. The random sample is augmented with some extra observations of a (presumably) rare choice.

3. $S = M$, $\mathcal{J}(1) = \{1\}$, $\mathcal{J}(2) = \{2\}$, $\ldots, \mathcal{J}(S) = \{M\}$. In this case predetermined proportions of the sample consist of each of the choices. It is called a *pure choice-based* sample. It simplifies notation but at the same time the distinction between stratum indicator s and choice i becomes blurred.

The joint density of (s, i, x) is the product of the marginal probability of s, H_s, and the conditional density of i and x given the stratum. The latter is

$$\frac{f(i,x)}{\sum_{i' \in \mathcal{J}(s)} \int_{\mathcal{X}} f(i',z)dz}$$

and the product can be written as:

$$(4) \quad g(s,i,x) = H_s \frac{P(i|x,\theta)r(x)}{\sum_{i' \in \mathcal{J}(s)} Q(i')}$$

This is the density function induced by the sampling scheme, as opposed to the density function in the population (1). As a rule $f(.)$ will denote population density and probability functions, and $g(.)$ density and probability functions induced by the sampling scheme. The latter will sometimes loosely be referred to as *sampling* densities.

The complications in estimation of choice-based sampling models arise because maximization of the log likelihood function corresponding to this

density is not possible without parametrizing the marginal density of x in the population, $r(x)$. If the sampling were random, and consequently the density of the data is (1), the maximization of the logarithm of the likelihood function is no problem. The density $r(x)$ disappears after taking derivatives with respect to θ. This can be extended to the case where the sampling depends on the regressors x. The density induced by the sampling would then be:

$$(5) \quad h(i,x) = P(i|x,\theta)q(x)$$

with $q(x) \neq r(x)$. In this *exogenous sampling* case there is still no problem in maximizing the logarithm of the likelihood function because the density of x still factors out.

To stress the reciprocal relation between H and Q we will also define $H(i)$ and Q_s:

$$(6) \quad Q_s = \sum_{i \in \mathcal{J}(s)} Q(i)$$

$$(7) \quad H(i) = Q(i) \sum_{s|i \in \mathcal{J}(s)} \frac{H_s}{Q_s}$$

If, for no s, $i \in \mathcal{J}(s)$, then $H(i) = 0$. $H(i)$ is the marginal probability of choice i induced by the choice–based sampling, or again somewhat loosely, the *sample probability* of choice i. It is not to be confused with the *sample frequency* of choice i, $\hat{H}(i) = \sum I[i_n = i]/N$. In the population the marginal probability of choice i is $Q(i)$, but the sampling scheme multiplies this by the sum of the bias factors H_s/Q_s. The essence of choice–based sampling is that for some i the distortion factor $\sum_{s|i \in \mathcal{J}(s)} H_s/Q_s$ differs from unity, or, equivalently, for some i, the population probability $Q(i)$ is not equal to the sample probability $H(i)$. The marginal probability that an observation randomly drawn from the population is in $\mathcal{J}(s)$ is Q_s. Note that while the $H(i)$, H_s and the $Q(i)$ add up to one, the sum of the Q_s does not have to add up to one.

The following two examples will be used throughout the paper to clarify concepts.

Example 1 *Consider a model with two choices $i = 0, 1$ and two samples $s = 1, 2$. With probability $H_1 = h$ an observation is drawn from $\mathcal{J}(1) = \{0\}$ and with probability $H_2 = 1 - h$ it is drawn from $\mathcal{J}(2) = \{1\}$. The population probability of choice $i = 0$ is $Q(0) = q$. The density of the data is*

$$g(s, i, x) = \Big[\frac{h}{q} P(0|x, \theta)\Big]^{1-i} \Big[\frac{1-h}{1-q} P(1|x, \theta)\Big]^{i} r(x)$$

This is an example of the *pure choice-based sampling* case. Each stratum corresponds exacty to one choice and the distinction between s and i becomes irrelevant. The density is written as a function of i and x alone but could also have been written as a function of s and x alone.

Example 2 *Consider the model in example 1 with an additional, third subsample that consists of a random sample of the whole population. So $S = 3$, $H_1 = h_1$, $H_2 = h_2$ and $H_3 = 1 - h_1 - h_2$. The subpopulations are $\mathcal{J}(1) = \{0\}$, $\mathcal{J}(2) = \{1\}$ and $\mathcal{J}(3) = \{0, 1\}$. The probabilities associated with them $Q_1 = Q(1) = q$, $Q_2 = Q(2) = 1 - q$ and $Q_3 = Q(1) + Q(2) = 1$. The density of the observations is in this case:*

$$g(s, i, x) = \Big[\frac{h_1}{q} P(0|x, \theta)\Big]^{1-i} \cdot \Big[\frac{h_2}{1-q} P(1|x, \theta)\Big]^{i} \cdot r(x) \quad \text{for } s = 1, 2$$

$$= (1 - h_1 - h_2) P(0|x, \theta)^{1-i} P(1|x, \theta)^{i} r(x) \quad \text{for } s = 3$$

Here the pure choice-based sample is augmented with a random sample of the whole population. Now the distinction between the stratum s and the choice i is a real one, as will become more apparent later. Note that in the second example the Q_s add up to 2.

In the following it will be assumed that the investigator has a sample of N observations. N_s will denote the number of observations from subsample s and $N(i)$ the number of observations with choice i. Most of this paper will deal with the case where i, s and x are all observed. Later other cases will briefly be discussed. In the remainder of this paper the following assumptions will be maintained throughout. Other assumptions will be introduced when necessary.

Assumption 2.1 *$x \in \mathcal{X}$, $\mathcal{X}$ a compact subset of $\Re^L$. $i \in C$, C a finite set with M elements $\theta^* \in \text{int}\Theta$, Θ a compact subset of $\Re^K$.*

Assumption 2.2 *$P(i|x,\theta)$ is a twice continuously differentiable function of θ, and P and its first two derivatives with respect to θ are continous in x for all $\theta \in \Theta$. $P(i|x,\theta) > 0$ for all $i \in C$, $x \in \mathcal{X}$ and θ in an open neighbourhood of θ^*.*

Assumption 2.3 *For all $(\theta, Q) \neq (\theta^*, Q^*)$, there is an $A \subset \mathcal{X}$, an $i \in C$ and an $s \in \{1, \ldots, S\}$ such that*

$$\frac{\int_A P(i|x,\theta)dR(x)}{\sum_{i' \in \mathcal{J}(s)} Q(i')} \neq \frac{\int_A P(i|x,\theta^*)dR(x)}{\sum_{i' \in \mathcal{J}(s)} Q^*(i')}$$

Sometimes the following, weaker version of this identification assumption will be used:

Assumption 2.4 *For all $\theta \neq \theta^*$ there is an $A \subset \mathcal{X}$, an $i \in C$ such that:*

$$\int_A P(i|x,\theta)dR(x) \neq \int_A P(i|x,\theta^*)dR(x)$$

The data collection mechanism is the distinguishing feature of choice–based sampling and therefore a few more remarks about it are warranted. In the model as it has been set up sofar, the indicator s of the stratum to which the observation belongs is a random variable. Hence, N_s, the total number of observations from sample s is also a random variable. In fact, it has mean $H_s^* \cdot N$ and variance $H_s^* \cdot (1 - H_s^*) \cdot N$. In practice N_s is often not a random variable but a number fixed by the investigator prior to the data collection. To apply large sample theory however, we need a model for the data that goes beyond the current N observations. This model is provided by the assumption that for all $n \neq n'$ the random variables s_n $s_{n'}$ that indicate the strata from which the observations are drawn, are independent and identically distributed. The alternative is to work with exact distributions and condition on the value of N_s. This is in practice impossible. The assumptions made here, on the other hand, are not very restrictive. We estimate the parameters of interest $(\theta,\ Q)$ jointly with the parameters (H) of the multinomial distribution of s, or we use restrictions on the latter parameters if information about them is available. It will turn out that information about H is relatively useless in any case, in the sense that the asymptotic covariance matrix of estimators of θ^* and Q^* is not affected by it.

2.2 The WESML Estimator

The Weighted Exogenous Sampling Maximum Likelihood Estimator has been proposed by Manski and Lerman [22]. It estimates the parameters of the conditional probability given knowledge of the marginal probabilities. It was the first estimator for this problem and it is still appreciated for its computational ease. In the original paper the estimator was introduced in the *pure choice-based sampling case*, where each stratum corresponds to exactly to one choice. There are different ways to extend it to the general sampling framework. Therefore I will first give the pure choice-based sampling case and then discuss the merits of various extensions.

Manski and Lerman propose maximizing the weighted log likelihood function

$$(8) \quad L_{ML}(\theta) = \sum_{n=1}^{N} \frac{Q^*(i_n)}{H^*(i_n)} \ln P(i_n|x_n,\theta)$$

The estimator can be interpreted as a method of moments estimator[2] with moment vector

$$(9) \quad \psi(\theta,i,x) = \frac{Q^*(i)}{H^*(i)} \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta)$$

Taking the expectation of this function evaluated at θ^* over the sample density

$$(10) \quad g(i,x) = \frac{H^*(i)}{Q^*(i)} P(i|x,\theta^*) r(x)$$

gives zero under assumptions 2.1–2.5 as can be checked easily. $g(i,x)$ in (10) is a special case of (4) with $M = S$, $S_i = 1$ for all $i \in C$, implying that there is a function $\mathcal{S}(i) : M \to \{1,2,\ldots,S\}$ satisfying $Q_{\mathcal{S}(i)} = Q(i)$ and $H_{\mathcal{S}(i)} = H(i)$ for all i.

The weights in the name WESML are the $Q^*(i)/H^*(i)$, often denoted by $w(i)$. Observations that are of a type that occur more often in the sample than in the population are given lower weight than the observations that are undersampled. This is similar to the way in which surveys are made representative for a larger population. There, groups that are under–

[2]for a brief description of method of moments estimation see the appendix

or over- represented have weights greater or smaller than unity associated with them to make the weighted sample resemble the population closer in characteristics.

Alternatively, one could write the weights in this case as Q_s^*/H_s^*. Strata that are over–represented have a relatively low weight.

The extension to general sampling schemes can be done in different ways. Consider the scores for the random sampling likelihood multiplied by weights $w(i,s)$ that can depend on both the choice i and the stratum s:

$$(11)\quad \psi(\theta,i,s,x) = \frac{w(i,s)}{P(i|x,\theta)}\frac{\partial P}{\partial \theta}(i|x,\theta)$$

Its expectation over the density induced by the sampling scheme, evaluated at the true parameter values is:

$$(12)\quad \mathrm{E_g}\psi(\theta^*,i,s,x) = \int_{\mathcal{X}} \sum_{i=1}^{M}\Big[\sum_{s|i\in\mathcal{J}(s)} \frac{w(i,s)\cdot H_s^*}{\sum_{i'\in\mathcal{J}(s)} Q^*(i')}\Big]\frac{\partial P}{\partial \theta}(i|x,\theta^*)r(x)dx$$

A sufficient condition for this expectation to be zero is that the expression between the square brackets is equal to a constant, independent of i:

$$(13)\quad \sum_{s|i\in\mathcal{J}(s)} \frac{H_s^*}{\sum_{i'\in\mathcal{J}(s)} Q^*(i')} w(i,s) = 1$$

where the constant is normalized to one. Solutions for $w(i,s)$ are numerous[3] :

1. $w(i,s) = Q_s^*/(H_s^* \cdot S_i)$
2. $w(i,s) = Q^*(i)/H^*(i)$
3. $w(i,s) = 1$ for $s = \mathcal{S}(i)$ and 0 elsewhere, where $\mathcal{S}(i) : C \to \{1,2,\ldots,S\}$ is an arbitrary function, satisfying $i \in \mathcal{J}(\mathcal{S}(i))$.

A special case of the latter occurs if one of the strata (say the first) consists of the whole choiceset C and only the observations of that stratum are given any weight: $\mathcal{J}(1) = C$, $\mathcal{S}(i) = 1\ \forall i$. One effectively throws away all

[3]If the sampling is purely choice–based, i.e. one stratum s per choice i, they all reduce to $w(i_n,s_n) = Q(i_n)/H(i_n)$

observations that are not from the random subsample. An easy, but clearly not very efficient, solution to the problem of choice-based sampling.

Within the class of weights defined by (13) we can search for the most efficient one. This turns out to be a weighting scheme that does not involve the stratum s.

Theorem 2.1 *No estimator $\hat{\theta}$ of θ^* defined by:*

$$\sum_{n=1}^{N} w(i_n, s_n) \frac{1}{P(i_n|x_n, \hat{\theta})} \frac{\partial P}{\partial \theta}(i_n|x_n, \hat{\theta}) = 0$$

with $w(.,.)$ in the class defined by (13) has a asymptotic covariance matrix smaller than the covariance matrix of the estimator that has weights $w(i,s) = Q^*(i)/H^*(i)$

Proof: see appendix

In principle more general weighting schemes are possible. The score for the random sampling likelihood has expectation zero conditional on x. Hence we can multiply any weight in the class defined by (13) by a function of x to obtain another set of weights that will give moments with zero expectation.

Example 1 (continued) *The weights for this sampling scheme are a function of i alone:*

$$w(0) = \frac{q}{h}$$

$$w(1) = \frac{1-q}{1-h}$$

Example 2 (continued) *Three of the possible weighting schemes are:*

$$w(0,1) = q/(2h_1)$$

$$w(1,2) = (1-q)/(2h_2)$$

$$w(0,3) = w(1,3) = 1/((2 \cdot (1 - h_1 - h_2))$$

or:

$$w(0,s) = \frac{q}{h_1 + q \cdot (1 - h_1 - h_2)} \quad \text{for } s = 1,2,3$$

$$w(1,s) = \frac{1-q}{h_2 + (1-q) \cdot (1 - h_1 - h_2)} \quad \text{for } s = 1,2,3$$

or:

$$w(0,1) = w(1,2) = 0$$

$$w(0,3) = w(1,3) = 1$$

In the second case of the second example the weights are the same as in the first example if the sample proportion there were equal to $h = h_1 + q \cdot (1 - h_1 - h_2)$. This is the most efficient of the three weighting schemes. It may come as a surprise that the most efficient weighting scheme in this class ignores some information. This may even lead one to believe that this particular information, namely the stratum from which an observation is drawn, is irrelevant. This is not true as will become more apparent later. To give some intuition for the fact that the knowledge of s does contain information consider the following random variable that can be defined in the second example:

$$\Big(q - I[i = 0]\Big) \cdot I[s = 3]$$

This compares the probability of a choice 0 observation in the third stratum with the occurence with such an observation. It has expectation zero and could be used as a specification test or to increase efficiency. It could not be used if one did not know the stratum from which an observation was drawn.

In their paper [21] Manski and Lerman do not discuss the role of H^* in great depth. Cosslett showed that there was a complication. If one uses $\hat{H}(j) = \sum_{n=1}^{N} I[i_n = j]/N$ instead of $H^*(j)$ in (8) one increases efficiency. The asymptotic covariance matrix given in [21] is that based on H^*. To get the covariance matrix for the more efficient estimator one should use GMM theory with the moments

$$\psi(H,\theta,i,s,x)_1 = \frac{Q^*(i)}{H(i)} \cdot \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta)$$

$$\psi(H,\theta,i,s,x)_{2j} = H(j) - I[i=j] \qquad j = 1,\ldots,M-1$$

In affect one augments the parameter space with the $M-1$ probabilities $H(j)$. Later this unintuitive result that using the probability limit of a parameter rather than the estimate itself reduces efficiency will be presented in a different light.

2.3 The CML Estimator

The Conditional Maximum Likelihood Estimator was proposed by Manski and McFadden [23] in their survey of estimation techniques for choice-based samples. Despite the fact that the information in pure choice-based samples is in the conditional distribution of the exogenous variables x given i, they look at the conditional distribution of i given x. In the population this is $P(i|x,\theta)$ but the biased sampling has changed this to:

$$(14)\quad g(i|x) = \frac{P(i|x,\theta^*)H^*(i)/Q^*(i)}{\sum_{j=1}^{M} P(j|x,\theta^*)H^*(j)/Q^*(j)}$$

The marginal density of x in the population, $r(x)$, has factored out, just as in the random sampling case. Part of the potential loss in efficiency stems from the treatment of x as exogenous[4] while there is no cut in the likelihood function. The parameter of interest, θ, enters the conditional distribution of i given x as well as the marginal distribution of x. In fact, the marginal density of x is

$$(15)\quad g(x) = \sum_{s=1}^{S} \frac{H_s}{Q_s} \sum_{j\in\mathcal{J}(s)} P(j|x,\theta)r(x)$$

and that clearly involves θ. Nevertheless, one can still base inference on the conditional likelihood function.

Manski and McFadden propose maximizing the conditional likelihood:

$$(16)\quad L_{MM}(\theta) = \sum_{n=1}^{N} \ln \frac{P(i_n|x_n,\theta)H^*(i_n)/Q^*(i_n)}{\sum_{j=1}^{M} P(j|x_n,\theta)H^*(j)/Q^*(j)}$$

[4]for a discussion of exogeneity see Engle, Hendry and Richard [10].

Again the estimator can be interpreted as a method of moments estimator. The associated score vector is

$$(17)\quad \psi(\theta,i,x) = \frac{\partial P}{\partial \theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} - \Big[\sum_{j=1}^{M}\frac{\partial P}{\partial \theta}(j|x,\theta)\frac{H^*(j)}{Q^*(j)}\Big]\Big/\Big[\sum_{j=1}^{M}P(j|x,\theta)\frac{H^*(j)}{Q^*(j)}\Big]$$

A different route can also bring one to this estimator. Later the particular form of the estimator arrived at via this route will prove useful in comparing the CML estimator and the new estimator. Consider the joint probability of s and i given x. It is equal to

$$(18)\quad g(i,s|x) = \frac{P(i|x,\theta)H_s^*/Q_s^*}{\sum_{s=1}^{S}\frac{H_s^*}{Q_s^*}\sum_{j\in\mathcal{J}(s)}P(j|x,\theta)}$$

The score vector associated with the conditional likelihood based on $g(s,i|x)$ is

$$(19)\quad \psi(s,i,x) = \frac{\partial P}{\partial \theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} - \Big[\sum_{s=1}^{S}\frac{H_s^*}{Q_s^*}\sum_{j\in\mathcal{J}(s)}\frac{\partial P}{\partial \theta}(j|x,\theta)\Big]\Big/\Big[\sum_{s=1}^{S}\frac{H_s^*}{Q_s^*}\sum_{j\in\mathcal{J}(s)}P(j|x,\theta)\Big]$$

It might seem at first that these estimators, the one with score vector (17) and the one with score vector (19), are very different. This is not the case. The ratio of (18) to the probability of i given x gives the probability of s given i and x. It can be written as:

$$(20)\quad g(s|i,x) = \frac{H_s}{Q_s}\Big/\Big[\sum_{t|i\in\mathcal{J}(t)}\frac{H_t}{Q_t}\Big]$$

If we do not know $Q(j)$ the parameters of this distribution would contain information. But the CML method is only applicable if we do know the marginal population probabilities and in that case the conditional distribution of s given i and x is uninformative. Therefore maximization of the conditional likelihood of s and i given x leads to exactly the same estimator as the one based on maximization of the conditional likelihood of i given x as given in (14).

Example 1 (continued) *The score vector associated with this sampling scheme and the CML estimator is*

$$\psi(\theta,i,x) = \frac{\partial P}{\partial \theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} - \frac{[h/q-(1-h)/(1-q)]\frac{\partial P}{\partial \theta}(0|x,\theta)}{P(0|x,\theta)h/q + P(1|x,\theta)(1-h)/(1-q)}$$

Example 2 (continued) *In this case the score vector is the same as for the first example with h replaced by $h_1 + q \cdot (1-h_1-h_2)$. It can also be based on $g(s,i|x)$ and in that case it would be written as:*

$$\psi(\theta,i,x) = \frac{\partial P}{\partial \theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} - \frac{[h_1/q-h_2/(1-q)]\frac{\partial P}{\partial \theta}(0|x,\theta)}{P(0|x,\theta)h_1/q + P(1|x,\theta)h_2/(1-q) + 1 - h_1 - h_2}$$

Again the randomness of the third subsample is not used. Note that the expectation of (17) is zero conditional on x. One could therefore multiply it by any function of x and the parameters to obtain another moment vector that could be used for estimation purposes (subject to regularity conditions as those on the Jacobian of the transformation). For the model of example 1 the moment vector for the WESML estimator could be obtained by multiplying the moment vector of the CML estimator by

$$\frac{H(0)\cdot H(1)}{[H(0)-Q(0)]\cdot P(0|x,\theta) - Q(0)\cdot H(1)}$$

Amemiya and Vuong [3] compare the asymptotic covariance matrices of the CML and WESML estimators. They find that CML is at least as efficient as WESML, i.e. the difference between the asymptotic covariance matrix of the WESML estimator and that of the CML estimator is a positive semi-definite matrix. However, Cosslett finds that both estimators can be improved upon by replacing the true parameter values H^* of the sampling design by their maximum likelihood estimates or sample frequencies $\hat{H}$. In the same way as was done for the WESML estimator in the last section, one can derive the asymptotic covariance matrix for the improved CML estimator by considering the GMM estimator for θ and H based on the moments

$$(21)\quad \psi(\theta,H,i,x)=\frac{\partial P}{\partial\theta}(i|x,\theta)\frac{1}{P(i|x,\theta)}$$
$$-\Big[\sum_{j=1}^{M}\frac{\partial P}{\partial\theta}(j|x,\theta)\frac{H(j)}{Q^{*}(j)}\Big]\Big/\Big[\sum_{j=1}^{M}P(j|x,\theta)\frac{H(j)}{Q^{*}(j)}\Big]$$

$$(22)\quad \psi_{2j}(\theta,H,i,x)=H(j)-I[i=j]$$

The modified estimators can in general not be ranked by comparing the asymptotic covariance matrices.

2.4 The PML Estimator

Cosslett [7,8,9] proposed the Pseudo Maximum Likelihood Estimator. Consider the likelihood function based on the density function (4). It cannot directly be maximized over the parameter space and the space of densities $r(x)$. If one replaces the density $r(x_n)$ by a set of discrete weights r_n, such that $\sum_{n=1}^{N} r_n = 1$ and $r_n \geq 0$, maximization is possible. One would obtain the following program:

$$(23)\quad \max_{\theta,r_n}\sum_{n=1}^{N}\ln\Big[\frac{H_{s_n}P(i_n|x_n,\theta)r_n}{\sum_{i'\in\mathcal{J}(s_n)}\sum_{n'=1}^{N}P(i'|x_{n'},\theta)r_{n'}}\Big]\quad\text{subject to}\quad\sum_{n=1}^{N}r_n=1$$

This is off course no longer a proper likelihood but at least it can be maximized. The solution of the maximization over r_n and θ turns out to be equivalent to the solution of the problem

$$(24)\quad \max_{\theta\in\Theta}\max_{\lambda\in\Lambda_1}\sum_{n=1}^{N}\ln\frac{\lambda(s_n)P(i_n|x_n,\theta)}{\sum_{s=1}^{S}\lambda(s)\sum_{j\in\mathcal{J}(s)}P(j|x_n,\theta)}$$

where

$$\Lambda_1=\Big\{\lambda\in R^S\Big|\lambda\geq 0,\sum_{n=1}^{N}\Big[\sum_{s=1}^{S}\lambda(s)\sum_{j\in\mathcal{J}(s)}P(j|x_n,\theta)\Big]=N\Big\}$$

λ has to be normalized in this maximization. The particular normalization chosen here will later facilitate comparisons with other estimators. This estimator does not require knowledge of Q^{*}. Note that H^{*} does not feature in it. The consistency of this estimator has to be proven directly. In the

interpretation as maximum likelihood estimator it has more parameters, $(N + K)$, than there are observations, (N). The nuisance parameters $\lambda(s)$ have probability limit H_s^*/Q_s^* for $s = 1, 2, \ldots, S$. The asymptotic properties for this estimator are most easily derived by writing it as a method of moments estimator. To do that with the particular nomalization chosen, it is convenient to add H as a parameter. As long as the $\hat{\lambda}$ and $\hat{\theta}$ maximizing (24) are interior solutions, they can also be characterized by $\sum_{n=1}^{N} \psi_{C1}(\hat{H}, \hat{\theta}, \hat{\lambda}, i_n, s_n, x_n) = 0$, with $\psi_{C1} = (\psi_1' \ \psi_2' \ \psi_3')'$, and

$$(25)\quad \psi_1(\lambda, \theta, s, i, x) = \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta)$$

$$- \Big[\sum_{t=1}^{S} \lambda(t) \sum_{j \in \mathcal{J}(t)} \frac{\partial P}{\partial \theta}(j|x,\theta)\Big] \Big/ \Big[\sum_{t=1}^{S} \lambda(t) \sum_{j \in \mathcal{J}(t)} P(j|x,\theta)\Big]$$

$$(26)\quad \psi_2(\lambda, \theta, H, s, i, x)_t =$$

$$\frac{H_t}{\lambda(t)} - \Big[\sum_{j \in \mathcal{J}(t)} P(j|x,\theta)\Big] \Big/ \Big[\sum_{s'=1}^{S} \lambda(s') \sum_{j \in \mathcal{J}(s')} P(j|x,\theta)\Big]$$

$$(27)\quad \psi_3(H, s)_t = I[s = t] - H_t$$

Cosslett proves that the estimator for θ^* is efficient in the class of asymptotically unbiased estimators.

In the case that Q^* is known, Cosslett proposes maximization of the same function, (23), under the restriction that for all $j \in C$, we have $Q(j) = \sum_{n=1}^{N} r_n P(j|x_n, \theta)$. This system is equivalent to

$$(28)\quad \max_{\theta \in \Theta} \min_{\lambda \in \Lambda_2} \sum_{n=1}^{N} \ln \frac{P(i_n|x_n,\theta)}{\sum_{j=1}^{M} \lambda(j) P(j|x_n,\theta)}$$

where

$$\Lambda_2 = \left\{ \lambda \in R^M \middle| \sum_{j=1}^{M} \lambda(j) Q(j) = 1, \sum_{j=1}^{M} \lambda(j) P(j|x_n,\theta) \geq 0 \ \forall n = 1, \ldots N \right\}$$

The same problems with proving consistency and asymptotic normality as above apply. Note that in contrast to the function maximized before, (24), this objective function does not depend on the stratum indicators s. Once the population proportions Q are known these do not contain information anymore. The dimension of λ has changed from S to M. The probability limit is in this case $H^*(i)/Q^*(i)$. The limit is the ratio of the sampling and population probabilities of the choices rather than of the strata as in the previous case. A method of moments representation of the estimator is possible with the moments $\psi_{C2} = (\psi_1' \ \psi_2')'$ defined by

$$(29)\quad \psi_1(\lambda,\theta,i,x) = \frac{1}{P(i|x,\theta)}\frac{\partial P}{\partial \theta}(i|x,\theta) - \Big[\sum_{j=1}^{M}\lambda(j)\frac{\partial P}{\partial \theta}(j|x,\theta)\Big]\Big/\Big[\sum_{j=1}^{M}\lambda(j)P(j|x,\theta)\Big]$$

$$(30)\quad \psi_2(\lambda,\theta,x)_j = \Big[P(j|x,\theta) - P(M|x,\theta)Q^*(j)/Q^*(M)\Big]\Big/\Big[\sum_{j'=1}^{M}\lambda(j')P(j'|x,\theta)\Big]$$

and $\lambda(M) = (1 - \sum_{j=1}^{M-1}\lambda(j)Q^*(j))/Q^*(M)$. If H^* were known, the probability limit of λ would also be known. Cosslett proves however that his estimator of θ^* is efficient, independent of the information on H^* available. This is only possible if asymptotically $\hat{\lambda}$ and $\hat{\theta}$ are uncorrelated, which in fact is the case. The computational difficulties stem from the very different nature of the parameters of the optimization program, λ and θ. Most optimization algorithms treat all parameters in the same way and in this case that does not work very well.

If one subsitutes the probability limit of λ into (28), and maximizes this function over θ one obtains the in general inefficient CML estimator. This could not happen if (28) were a proper likelihood function with parameters λ and θ. It does suggest a way though to reduce the dimensionality of the computational problem of solving $\sum_{n=1}^{N}\psi_{C2}(\hat{\lambda},\hat{\theta},i_n,s_n,x_n) = 0$ without losing efficiency. One could add the moment (30), evaluated at the probability limit of λ, H^*/Q^*, to the score for the CML likelihood, (17) to get an efficient method of moments estimator.

One would still be left with completely separate estimation procedures for the case with known and the case with unknown Q. This cannot be remedied by using the method of moments estimator based on the moments (25), (26) and (27) with H_s^*/Q_s^* substituted in for $\lambda(s)$. That would give an inefficient estimator for θ.

Example 1 (continued) *the function to be maximized in Cosslett's procedure for the unknown Q model is:*

$$L_{C1}(\theta,\lambda) = \ln \sum_{n=1}^{N} \frac{\left[P(0|x_n,\theta)\lambda(1)\right]^{1-i_n} \cdot \left[P(1|x_n,\theta)\lambda(2)\right]^{i_n}}{\lambda(1)P(0|x_n,\theta) + \lambda(2)P(1|x_n,\theta)}$$

for the known Q model the objective function is:

$$L_{C2}(\theta,\lambda) = \ln \sum_{n=1}^{N} \frac{P(0|x_n,\theta)^{1-i_n} P(1|x_n,\theta)^{i_n}}{\lambda(1)P(0|x_n,\theta) + \lambda(2)P(1|x_n)}$$

Example 2 (continued) *The objective function for the unknown Q case is for this model*

$$L_{C1}(\theta,\lambda) = \ln \sum_{n=1}^{N} \frac{P(0|x_n,\theta)^{1-i_n} P(1|x,\theta)^{i_n} \lambda(s_n)}{\lambda(1)P(0|x_n,\theta) + \lambda(2)P(1|x_n,\theta) + \lambda(3)}$$

and for the known Q case:

$$L_{C2}(\theta,\lambda) = \ln \sum_{n=1}^{N} \frac{P(0|x_n,\theta)^{1-i_n} P(1|x_n,\theta)^{i_n}}{\lambda(1)P(0|x_n,\theta) + \lambda(2)P(1|x_n,\theta)}$$

Notice that for the known Q case the two sampling schemes give exactly the same estimators. The fact that there are different strata does not affect the form of the estimator once the marginal choice probabilities are known. In this context it is worth noting that if $Q(0)$ is not known, it would be identified non–parametrically in the second example but not in the first. In the second example a non–parametric estimator for $Q(0)$ would be $\sum_{n=1}^{N} I[s_n = 3] \cdot I[i_n = 0] \big/ \sum_{n=1}^{N} I[s_n = 3]$. It is therefore clear that the estimation problems are very different for the two examples in the case that the population shares are not known.

3 An efficient GMM Estimator for Choice–based Samples

In this section the new estimator will be discussed. The strategy is as follows: first it will be assumed that the regressors x have a discrete distribution with known points of support. This is off course very restrictive but it enables one to use standard maximum likelihood theory. In particular the Cramér–Rao bound can be calculated and used as an efficiency bound. Potential restrictions in the form of knowledge of the marginal probabilities can easily be incorporated in this case.

The maximum likelihood estimator for the discrete regressor case can be written in such a way that the knowledge of the points of support is not used explicitly. It turns out that the estimator remains valid even if the distribution of x is continous. Efficiency will be proven for this estimator in the general case. The theory behind the Cramér–Rao bound is no longer applicable and therefore a generalization of this concept from Hájek [13], used in the econometric literature by Chamberlain [5,6], will be applied. One difference with Chamberlains results deserves mention. Chamberlain obtains the result that if one has conditional moment restrictions, one has to increase the number of moments with the sample size to reach efficiency. Here we do have conditional moment restrictions but the number of moments needed for efficiency is fixed (it will turn out to be $K + M + S - 1$). The intuition is that because those moments that are conditional restrictions, are derived as scores to the conditional likelihood function, they contain all the information that is in the conditional model.

To give some intuition for the way in which assuming a discrete distribution can lead one to estimators that are valid and efficient even if the distribution is in fact continuous, consider the following example. It is similar to one in Chamberlain [5]. Suppose one is interested in the probability that a random variable z is positive, $\delta = \Pr(z > 0)$. If z is known to have a discrete distribution with points of support $\{z^1, z^2, \ldots, z^L\}$, and with unknown probabilities $\{p_1, \ldots, p_L\}$, one could estimate δ on the basis of N independent observations $\{z_1, z_2, \ldots, z_N\}$ by maximum likelihood techniques as:

$$\hat{\delta} = \sum_{m|z^m>0} \hat{p}_m = \sum_{m|z^m>0} \frac{1}{N} \sum_{n=1}^{N} I[z_n = z^m] = \frac{1}{N} \sum_{n=1}^{N} I[z_n > 0]$$

In the last representation of the estimator it does not depend explicitly on the points of support, only on the realized observations. It can also be used as an estimator for δ if z does not have a discrete distribution. In fact, whatever the distribution of z, $\hat{\delta}$ is a very good estimator, and efficient in a sense to be defined later.

3.1 The Case with Discrete Exogenous Variables

The subject of this section is the case where x has a discrete distribution. This will allow one to use standard Maximum Likelihood theory. Few formal proofs of consistency and asymptotic properties of estimators will be given in this section. The main point here, as indicated earlier in the introduction to section 3, is to use Maximum Likelihood theory to guide one to an estimator that will be used outside the Maximum Likelihood framework.

Assumption 3.1 *x is a discrete random variable with probability $\pi_m > 0$ at x^m for $m = 1, 2, \ldots, L$, and the masspoints x^m, elements of an Euclidean space, are known. L is larger than M.*

An observation (s, i, x) can now be written as (s, i, l), where l indicates the x type of the observation. The log likelihood function for the observations $(s_n, i_n, l_n)_{n=1}^N$ is:

$$(31) \quad L(H, \pi, \theta) = \sum_{n=1}^{N} \ln H_{s_n} + \ln P(i_n | x^{l_n}, \theta) + \ln \pi_{l_n}$$

$$- \ln \sum_{j \in \mathcal{J}(s_n)} \sum_{m=1}^{L} \pi_m P(j | x^m, \theta)$$

π_L is shorthand for $1 - \sum_{l=1}^{L-1} \pi_l$. The likelihood equations corresponding to this problem are:

$$(32) \quad \frac{\partial L}{\partial H_t}(H, \pi, \theta) = \sum_{n=1}^{N} \frac{I[s_n = t]}{H_t} - \frac{I[s_n = S]}{H_S}$$

$$(33)\quad \frac{\partial L}{\partial \pi_m}(H,\pi,\theta) = \sum_{n=1}^{N} \frac{I[x^{l_n} = x^m]}{\pi_m} - \frac{I[x^{l_n} = x^L]}{\pi_L}$$

$$+ \Big[\sum_{j\in\mathcal{J}(s_n)} P(j|x^L,\theta) - P(j|x^m,\theta)\Big] \Big/ \Big[\sum_{j\in\mathcal{J}(s_n)} \sum_{m'=1}^{L} \pi_{m'} P(j|x^{m'},\theta)\Big]$$

$$(34)\quad \frac{\partial L}{\partial \theta}(H,\pi,\theta) = \sum_{n=1}^{N} \frac{\partial P}{\partial \theta}(i_n|x^{l_n},\theta) \frac{1}{P(i_n|x^{l_n},\theta)}$$

$$- \Big[\sum_{j\in\mathcal{J}(s_n)} \sum_{m=1}^{L} \pi_m \frac{\partial P}{\partial \theta}(j|x^m,\theta)\Big] \Big/ \Big[\sum_{j\in\mathcal{J}(s_n)} \sum_{m=1}^{L} \pi_m P(j|x^m,\theta)\Big]$$

Let $\hat{H}$, $\hat{\pi}$ and $\hat{\theta}$ be the maximum likelihood estimators and assume that they are the unique solutions to the system obtained by setting the likelihood equations equal to zero. In this discrete regressor framework the maximum likelihood estimator for $Q(j)$ is:

$$(35)\quad \hat{Q}(j) = \sum_{l=1}^{L} \hat{\pi}_l P(j|x^l,\hat{\theta})$$

We can draw a few conclusions from these scores. The fact that the derivatives $\partial L/\partial\theta$ and $\partial L/\partial\pi$ do not depend on H implies that the asymptotic covariance matrix has a block diagonal structure. Asymptotically $\hat{H}$ and $\hat{\theta}$ are uncorrelated and knowledge of H does not enhance our ability to estimate θ, π or functions thereof. This is not surprising and it holds for other estimators than the above. Independent of the amount of information one has about the functions $P(j|x)$ and the vector Q, H^* will optimally be estimated by $\hat{H}$, statisfying $\hat{H}_t = \sum_{n=1}^{N} I[s_n = t]/N$. The covariance matrix of $\sqrt{N}(\hat{H} - H^*)$ only depends on H^*. If the covariance of $\hat{H}$ and any estimator $\tilde{\theta}$ of θ^* and $\tilde{Q}$ of Q^* were non-zero, knowledge of θ^* and Q^* would reduce this variance. That is impossible since it is the variance that would apply even if one knew the functions $P(j|x)$ and Q exactly. Therefore, $\hat{H}$ is independent of any estimator of θ^* and Q^* and knowledge of H^* can never help one to estimate them more accurately. However, it will be convenient to treat H the same way as the other parameters θ and Q. One should

bear in mind that even if H_s were known, there would be no harm in doing so.

The next step is to transform the parameter vector into one that includes Q. This serves three purposes. Firstly, the system of equations describing the maximum likelihood estimators in the transformed model has a recursive structure. This will imply that in order to calculate $\hat{\theta}$ one has to solve a system of $K + M - 1$ equations. This is a significant reduction from the $K + L - 1$ dimensional system that has to be solved to obtain $\hat{\theta}$ at this moment, assuming that the number of points of support of x, L, is much larger than the number of choices, M. Secondly, it will provide an easier framework for analyzing estimation with restrictions on Q. In the transformed model it will be a conventional maximum likelihood estimation problem with linear restrictions on the parameters. Thirdly, and most importantly, the estimators for θ^* and Q^* can, after the transformation, be written in a form that does not require knowledge of the points of support.

Define the $(M-1) \times L$ dimensional matrix V to be the matrix with typical element

$$(36) \quad v_{il} = P(i|x^l, \theta^*)$$

Partition V into $(V_0 \; V_1)$ with V_0 a square matrix. The condition that will allow us to do the desired transformation is that V_0 is non–singular, possibly after reordering the points of support. Assume that this condition is satisfied. Partition π into $(\pi_1 \; \pi_2)$ with $\dim(\pi_1) = M - 1$ and $\dim(\pi_2) = L - M$. The Jacobian of the transformation from the vector $(H \; \theta \; \pi_1 \; \pi_2)$ to $(H \; Q \; \theta \; \pi_2)$ is non–zero as a consequence of the above condition. The maximum likelihood estimators for the new parameter vector can be written as

$$(37) \quad \sum_{n=1}^{N} \psi(\hat{H}, \hat{\theta}, \hat{\pi}_2, \hat{Q}, i_n, s_n, l_n) = 0$$

where $\psi = (\psi_1' \; \psi_2' \; \psi_3' \; \psi_4')'$ with ψ_1 an $S - 1$ vector, ψ_2 an $M - 1$ vector, ψ_3 a K vector and ψ_4 an $L - M$ vector with typical elements:

$$(38) \quad \psi_1(H, \theta, \pi_2, Q, i, s, l)_t = H_t - I[s = t]$$

$$(39) \quad \psi_2(H, \theta, \pi_2, Q, i, s, l)_j = Q(j) - P(j|x^l, \theta) \Big/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} P(i'|x^l, \theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')} \right]$$

$$(40)\quad \psi_3(H,\theta,\pi_2,Q,i,s,l) = \frac{\partial P}{\partial \theta}(i|x^l,\theta)\frac{1}{P(i|x^l,\theta)}$$

$$- \Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)} \frac{\partial P}{\partial \theta}(i'|x^l,\theta)}{\sum_{i'\in\mathcal{J}(t)} Q(i')}\Big] \Big/ \Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)} P(i'|x^l,\theta)}{\sum_{i'\in\mathcal{J}(t)} Q(i')}\Big]$$

$$(41)\quad \psi_4(H,\theta,\pi_2,Q,i,s,l)_m = \pi_{2m} - I[x^l = x^m] \Big/ \Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)} P(i'|x^l,\theta)}{\sum_{i'\in\mathcal{J}(t)} Q(i')}\Big]$$

The first three parts of the ψ vector do not depend on π_2. They can therefore be solved seperately as a function of H, Q and θ. Since the solution for $\hat{H}$ is trivial, the system that has to be solved to obtain $\hat{\theta}$ is reduced to a $K+M-1$ dimensional one. Note that the only way in which the moments (38)-(40) depend on the mass points is via the observed x values. This is very similar to the example in the introduction to section 3. It implies that the Maximum Likelihood estimators for Q, H and θ can be calculated without knowing apriori what the masspoints of the random variable x are. In fact it will be seen in section 3.1 that one does not even need the assumption that x has finite support.

The three moments have clear interpretations. When evaluated at $Q = Q^*$ and $H = H^*$, the third moment ψ_3 is equal to the score for the conditional likelihood of i and s given x. Compare (40) with (19). If the sampling scheme were random (say $S = 1$ and $\mathcal{J}(1) = C$) the second moment would compare the marginal probability with the average of the conditional probabilities. The choice-based sampling scheme implies that before the comparison can be made the conditional probabilities have to be weighted to correct for the sampling induced bias. The first moment, ψ_1 is easy to interpret but it is difficult to explain why it has to be in the moment vector. The importance is clear from Cosslett's [7,8,9] result that using sample frequencies instead of the true H^* in the WESML and CML estimators increases efficiency. It is also clear that using optimal method of moments estimation with $\psi(H^*,\theta,Q,\pi_2)$ is at least as efficient as the estimator defined above (which is the method of moments estimator with moments $\psi(H,\theta,Q,\pi_2)$). For the moment the explanation will be left open. It is easier to discuss in an explicit method of moments framework rather than the maximum likelihood framework we are currently using.

The other advantage of the transformation alluded to earlier is the ease with which information about Q can be incorporated. Before the transformation this would have amounted to a maximization in $K+L+S-2$ space with $M-1$ restrictions. Now it will turn out to involve a maximization in $K+S-2$ space. The following lemma gives an efficient way of using restrictions on some parameters if one has the recursive structure we have derived above. Note that the structure is very similar to that analyzed by Newey [24] in his discussion of sequential estimators.

Lemma 3.1 *Suppose the maximum likelihood estimator of a vector β with $\beta = (\beta_1' \ \beta_2' \ \beta_3')'$ can be characterized by*

$$\sum_{n=1}^{N} h_1(\hat{\beta}_1, \hat{\beta}_2, x_n) = 0$$

$$\sum_{n=1}^{N} h_2(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, x_n) = 0$$

with $dim(h_1) = dim(\beta_1) + dim(\beta_2)$ and $dim(h_2) = dim(\beta_3)$. Then, the optimal constrained method of moments estimator for β_1 given $\beta_2 = 0$ based on minimization of

$$\frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1, 0, x_n)' \cdot A_N \cdot \frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1, 0, x_n)$$

where $A_N \xrightarrow{a.s.} Eh_1 \cdot h_1'$, has the same asymptotic covariance matrix as the constrained maximum likelihood estimator. In other words, it achieves the Cramér-Rao lower bound.

proof: see appendix

The relevance for the problem analyzed in this section is clear. If one has linear restrictions of the form[5] $b_1 Q + b_2 H + b_3 \theta = b_0$, and if one is only interested in estimation of θ^*, Q^* or H^* or a subset thereof, one does not need to resort to maximizing the constrained likelihood function. It is as

[5] In practice the most useful restriction included in this class is $Q = Q^*$, the case analyzed in great detail by Manski and Lerman[22], Manski and McFadden[23] and by Cosslett [7] as a special case

efficient asymptotically, in the sense of the covariance matrix, to estimate these parameters with the method of moments, using (38)–(40) as moments. One provision is that because there are more moments than free parameters, that is, because there are binding restrictions, one has to weigh the moments optimally.

Now the relevance of the first moment, ψ_1 in (38), or equivalently the issue of using $\hat{H}$ or H^* can be studied more conveniently. Define the following method of moments estimators[6] all conditional on the true value for Q.

$\hat{\theta}^1$ the method of moments estimator for θ^* using $\psi_3(H^*, \theta, Q^*, i, s, l)$ as moments. This is the CML estimator as proposed by Manski and McFadden [22].

$\hat{\theta}^2$ the method of moments estimator for θ^* using $\psi_3(H^*, \theta, Q^*, i, s, l)$ and $\psi_1(H^*, \theta, Q^*, i, s, l)$ as moments

$\hat{\theta}^3$ the method of moments estimator for θ^* based on joint estimation with H, using $\psi_1(H, \theta, Q^*, i, s, l)$ and $\psi_3(H, \theta, Q^*, i, s, l)$ as moments. This is the improved version of the CML estimator as proposed by Cosslett.

Using the asymptotic covariance matrix as the criterium, one can rank these estimators. Use the notation $\hat{\theta}^j \succeq \hat{\theta}^i$ if the difference between the covariance matrix of $\hat{\theta}^j$ and the covariance matrix of $\hat{\theta}^i$ is a positive semi-definite matrix, $\hat{\theta}^j \succ \hat{\theta}^i$ if the difference is positive definite and $\hat{\theta}^j \sim \hat{\theta}^i$ if the asymptotic covariance matrices are equal. $\hat{\theta}^2 \succeq \hat{\theta}^1$ because $\hat{\theta}^2$ uses more moments for the same parameters. $\hat{\theta}^2 \succeq \hat{\theta}^3$ because $\hat{\theta}^2$ uses the same moments but estimates fewer parameters. However, since knowledge of H^* was proven to be of no value in estimating θ^*, $\hat{\theta}^3 \sim \hat{\theta}^2$. Therefore $\hat{\theta}^3 \succeq \hat{\theta}^1$. The issue now is why $\hat{\theta}^2 \succ \hat{\theta}^1$ and $\hat{\theta}^3 \succ \hat{\theta}^1$ in general. An example from SUR (Seemingly Unrelated Regression) might clarify that. Consider the problem of estimating one parameter (α) on the basis of observations $(y_n, \varepsilon_n)_{n=1}^N$, with the following structure:

[6] with the method of moments estimator for β^* on the basis of the moments $h(z, \beta)$ we mean the minimand of a quadratic form $\frac{1}{N}\sum_{n=1}^N h(z_n, \beta)' \cdot C_N \cdot \frac{1}{N}\sum_{n=1}^N h(z_n, \beta)$ where $C_N \xrightarrow{a.s.} Eh(z, \beta^*) \cdot h(z, \beta^*)'$

$$\mathrm{E}\begin{pmatrix} y-\alpha \\ \varepsilon \end{pmatrix} = 0$$

$$\mathrm{E}\begin{pmatrix} y-\alpha & \varepsilon \end{pmatrix}\begin{pmatrix} y-\alpha \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

The variance of $\sqrt{N}(\hat{\alpha}-\alpha)$ based on the single moment $y-\alpha$ is 1. If both moments are used, this can be reduced to $1-\rho^2$, despite the fact that the additional moment does not contain any unknown parameters. The effect comes purely from the correlation of the moments. In the problem under consideration the extra moment $H_t - I[s_n = t]$ might add efficiency via the correlation with the other moments.

3.2 The General case

In the previous section it was assumed that x had a discrete distribution with known, finite support. In that case the maximum likelihood estimators for θ^*, H^*, Q^* and π were derived. It turned out that the estimators for the parameters of interest, θ^* and Q^* could be calculated by solving a smaller set of equations that did not involve π. In this section it will be shown that these equations can be used to give an efficient estimator even if x is not a discrete random variable. Assumption 3.1 will be replaced by the following:

Assumption 3.2 *x is a random variable with distribution function $R(x)$ and bounded support $\mathcal{X}$.*

The typical observation is now the triple $(s,i,x) \in \{1,2,\ldots,S\} \times C \times \mathcal{X}$. The first step is to rewrite the moments (38)-(40) slightly. Define $\psi = (\psi_1' \ \psi_2' \ \psi_3')'$, with ψ_1 an $S-1$ vector, ψ_2 an $M-1$ vector and ψ_3 a K vector with typical elements:

$$(42)\quad \psi_1(H,\theta,Q,i,s,x)_t = H_t - I[s=t]$$

$$(43)\quad \psi_2(H,\theta,Q,i,s,x)_j = Q(j) - P(j|x,\theta)\Big/\left[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)} P(i'|x,\theta)}{\sum_{i'\in\mathcal{J}(t)} Q(i')}\right]$$

$$(44)\quad \psi_3(H,\theta,Q,i,s,x) = \frac{\partial P}{\partial \theta}(i|x,\theta)\frac{1}{P(i|x,\theta)}$$

$$- \Big[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} \frac{\partial P}{\partial \theta}(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')}\Big] \Big/ \Big[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} P(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')}\Big]$$

In section 3.1 these moments were derived from likelihood equations. Therefore it was immediate that they had expectation zero. Here their validity as moments suitable for usage in a method of moments procedure has to be established directly. For all three of them it is easy to check that the expectation over the distribution induced by the sampling scheme, (for good order, $g(s,i,x)$ in (4)) is zero.

With these moment equations and a possibly stochastic weight matrix C_N the objective function $\mathrm{R}_N(\theta, Q, H)$ can be defined as:

$$(45) \quad \frac{1}{N}\sum_{n=1}^{N} \psi(H,\theta,Q,i_n,s_n,x_n)' \cdot C_N \cdot \frac{1}{N}\sum_{n=1}^{N} \psi(H,\theta,Q,i_n,s_n,x_n)$$

We will use the following shorthand: $\gamma = (H'\ \theta'\ Q')'$ and γ^* accordingly. Define:

$$\Delta_0 = E\ \psi(H^*,\theta^*,Q^*,i,s,x) \cdot \psi(H^*,\theta^*,Q^*,i,s,x)'$$

and

$$\Gamma_0 = E\frac{\partial \psi(H^*,\theta^*,Q^*,i,s,x)}{\partial (H'\ \theta'\ Q')}$$

Assumption 3.3

1. Δ_0 *is non-singular*
2. Γ_0 *has full rank* $(= K + M + S - 2)$

If assumption 3.3.1 is not fulfilled one should leave out some of the moments. Some of them are perfectly correlated and some therefore do not contribute any information. If the other assumption does not hold then asymptotic normality will be a problem. This is a rare problem though. Note that identification is already guaranteed by the assumptions made in section two.

The estimator $\hat{\gamma}$ of γ^* is defined as the minimand of $\mathrm{R}_N(\gamma)$ over the Cartesian product of the sets $\{H \in \Re^{S-1} | 0 \leq H_s \leq 1,\ 0 \leq \sum_{s=1}^{S-1} H_s \leq 1\}$, $\{Q \in \Re^{M-1} | 0 \leq Q(j) \leq 1,\ 0 \leq \sum_{j=1}^{M-1} Q(j) \leq 1\}$ and Θ. The following theorem gives its properties.

Theorem 3.2 *suppose that assumptions 2.1–2.4 and 3.2–3.3 hold. Then the estimator $\hat{\gamma}$ for γ^* converges almost surely to γ^* and satisfies:*

$$\sqrt{N}(\hat{\gamma} - \gamma^*) \xrightarrow{d} \mathcal{N}\left(0, \Gamma_0^{-1}\Delta_0\Gamma_0'^{-1}\right)$$

If we partition γ and Γ_0 in

$$\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \quad \Gamma_0 = \begin{pmatrix} \Gamma_{01} & \Gamma_{02} \end{pmatrix}$$

then we can estimate γ_1^* in the case γ_2^* is known with the minimand $\tilde{\gamma}_1$ of $R_N(\gamma_1, \gamma_2^*)$. $\tilde{\gamma}_1$ converges almost surely to γ_1^* and it satisfies:

$$\sqrt{N}(\tilde{\gamma}_1 - \gamma_1^*) \xrightarrow{d} \mathcal{N}\left(0, (\Gamma_{01}'C_0\Gamma_{01})^{-1}\Gamma_{01}'C_0\Delta_0C_0\Gamma_{01}(\Gamma_{01}'C_0\Gamma_{01})^{-1}\right)$$

proof: see appendix

The optimal method of moments estimator is the one with C_0, the limit of the weight matrix equal to Δ_0^{-1}. In that case the covariance matrix reduces to $(\Gamma_{01}'\Delta_0^{-1}\Gamma_{01})^{-1}$ for the restricted case. It is this estimator that will be analyzed as a candidate for efficiency.

Example 1 (continued) *The moment equations for this sampling scheme for the new estimator are:*

$$\psi_1(h, \theta, q, s, i, x) = h - I[s = 1]$$

$$\psi_2(h, \theta, q, s, i, x) = q - \frac{P(0|x, \theta)}{P(0|x, \theta)h/q + P(1|x, \theta)(1 - h)/(1 - q)}$$

$$\psi_3(h, \theta, q, s, i, x) = \frac{\partial P}{\partial \theta}(i|x, \theta)\frac{1}{P(i|x, \theta)}$$

$$- \frac{\partial P}{\partial \theta}(i|x, \theta)\frac{h/q - (1 - h)/(1 - q)}{P(0|x, \theta)h/q + P(1|x, \theta)(1 - h)/(1 - q)}$$

Example 2 (continued) *The moment equations for the sampling scheme of this example are:*

$$\psi_1(h_1, h_2, \theta, q, s, i, x)_1 = h_1 - I[s = 1]$$

$$\psi_1(h_1, h_2, \theta, q, s, i, x)_2 = h_2 - I[s = 2]$$

$$\psi_2(h_1, h_2, \theta, q, s, i, x) = q - \frac{P(0|x, \theta)}{P(0|x, \theta)h_1/q + P(1|x, \theta)h_2/(1-q) + 1}$$

$$\psi_3(h_1, h_2, \theta, q, s, i, x) = \frac{\partial P}{\partial \theta}(i|x, \theta)\frac{1}{P(i|x, \theta)}$$

$$- \frac{\partial P}{\partial \theta}(i|x, \theta)\frac{h_1/q - h_2/(1-q)}{P(0|x, \theta)h_1/q + P(1|x, \theta)h_2/(1-q) + 1}$$

The difference between the two sampling schemes in these examples is the additional moment equation ψ_{12}. The form of the moment equations does not change with the potential restrictions on the parameters as was the case with the estimators proposed by Cosslett.

In the last section it was shown that the estimator achieved the Rao–Cramer lower bound. Here, we are not in a Maximum Likelihood framework so we cannot use this bound directly. Instead, we will use a efficiency concept from Hajek [13], used by Chamberlain [5,6] to prove efficiency of Method of Moment estimators. The idea behind this Local Asymptotic Minimax concept is that we look at the expected loss for a particular estimator while letting the true value of the parameter vary over a small neighbourhood. An estimator is efficient in this sense if there is no estimator that does better everywhere in this neighbourhood. In this particular case we let, as did Chamberlain, also the distribution of x vary over neighbourhoods that will be defined shortly. Then it will be shown that no estimator does better than the one defined in theorem 3.2 in the neighbourhood of the true distribution of x and the true parameter values of θ and Q. In the appendix the efficiency concept and its relevance will be discussed in greater depth.

Let Π denote the space of probability measures over the set $\mathcal{Z}$. Then a neighbourhood F_ϵ of a measure F is:

$$(46) \quad \left\{ G \in \Pi \middle| \| \int b_j dG - \int b_j dF \| < \epsilon_j, j = 1, \ldots, K \right\}$$

for some continuous functions b_j with $\int \|b_j\| dF < \infty$. In words, two distributions are close to each other if a number of predetermined moments are close in absolute value.

The next step is to define the class of loss functions considered. If the method were very sensitive to the particular loss function used it would of course be of less interest. Fortunately this turns out not to be the case. A loss function $\ell : \Re \to \Re$ is an element of the class of loss functions $\mathcal{L}$ if:

1. for all $u \in \Re$ $\ell(u) = \ell(\|u\|)$
2. for all $u, v \in \Re$ $\|u\| \geq \|v\|$ implies $\ell(u) \geq \ell(v)$
3. $\int_{-\infty}^{\infty} \ell(u) \exp(-\lambda u^2/2) du < \infty$ for $\lambda > 0$
4. $\ell(0) = 0$

The result that we are interested in can now be stated:

Theorem 3.3 *For any estimator T_N of γ_1^*, any set of functions b_j, $j = 1, \ldots, K$ that define neighbourhoods, and for any loss function $\ell \in \mathcal{L}$, the expected minimax loss*

$$\lim_{\epsilon \downarrow 0} \lim_{N \to \infty} \inf_{T_N} \sup_{G \in R_\epsilon,\ \|\gamma - \gamma^*\| < \epsilon} E_{G\ \gamma}\, \ell(\sqrt{N}(T_N - \gamma_1))$$

$$\geq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \ell(\sigma u) \exp(-u^2/2) du$$

where σ is the square root of the (1,1) element of the covariance matrix of the optimal method of moments estimator. In other words, no estimator has lower expected[7] risk over a neighbourhood of the distribution of x and the parameters than the estimator in theorem 3.2

The formal proof will be given in the appendix but some intuition for the result will be presented here. For any continuous distribution over a compact set one can find discrete distribution with support in that compact set that has a predetermined set of moments in common with the continuous distribution. For these moments we chose the b_j that define the neighbourhoods, the moments ψ that are used in the estimation, and their outer products $\psi\psi'$ and derivatives $\frac{\partial \psi}{\partial \theta}$. Then we have a discrete distribution G in

[7] the expectation is taken over the distribution characterized by parameter γ and distribution G of the regressors x

the neighbourhood R_ϵ of the continuous distribution $R(x)$ that we started off with. Hence the bound on the continuous model cannot be lower than the one calculated for the discrete model. The bound for the discrete model is equal to the Rao–Cramer bound. Because the proposed estimator for the continuous model reaches this bound it must be efficient.

As in the discrete case, it does not matter whether we apply this to the estimation of the full vector $\gamma = (H' \ \theta' \ Q')'$ or to the estimation of γ_1 given $\gamma_2 = \gamma_2^*$. In both cases the bound on the loss is the loss for the (constrained) method of moments estimator.

In this method of moments framework it is easy to see how the restrictions on the marginal probabilities can be tested. For a general discussion of tests of this type see Newey [25]. In this particular case one would look at the value of $N \cdot \mathrm{R}_N(\hat{\theta}, Q^*, \hat{H})$ for a sequence of C_N converging to Δ_0^{-1}. Then:

$$(47) \quad N \cdot \mathrm{R}_N(\hat{\theta}, Q^*, \hat{H}) \xrightarrow{d} \mathcal{X}^2_{M-1}$$

These tests could be used to compare logit and probit specifications. Tests on H are not relevant since asymptotically the estimators for H and those for θ and Q are independent.

If the sampling were random these tests could still be employed. One of the ways to describe random sampling is $S = 1$, $\mathcal{J}(1) = C$. The the following $K + M - 1$ moments would be sufficient to estimate θ^* and Q^* or to test restrictions, $\psi = (\psi_2' \ \psi_3')'$:

$$(48) \quad \psi_2(\theta, Q, i, x)_j = Q(j) - P(j|x, \theta)$$

$$(49) \quad \psi_3(\theta, Q, i, x) = \frac{1}{P(i|x, \theta)} \frac{\partial P}{\partial \theta}(i|x, \theta)$$

Again these could be used to distinguish between logit and probit models.

3.3 The Connection with Cosslett's Estimators

The connection between the estimator proposed in the previous section and those proposed by Cosslett can best be seen by comparing the relevant moment vectors. In this section we will only show that Cosslett's estimator does not do better than the new one. First consider the case with known

Q. The moment vector for Cosslett's estimator is given in (29) and (30). It was argued there that $\lambda(j)$ could be replaced by its probability limit $H^*(j)/Q^*(j)$ without changing the asymptotic covariance matrix of $\hat{\theta}$. The moment vector would then be $\tilde{\psi} = (\tilde{\psi}_1' \ \tilde{\psi}_2')'$:

$$(50)\quad \tilde{\psi}_1(\theta, H^*, Q^*, s, i, x) = \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta)$$

$$\left[\sum_{j=1}^{M} \frac{\partial P}{\partial \theta}(j|x,\theta) H^*(j)/Q^*(j)\right] \Big/ \left[\sum_{j=1}^{M} P(j|x,\theta) H^*(j)/Q^*(j)\right]$$

$$(51)\quad \tilde{\psi}_2(\theta, H^*, Q^*, s, i, x)_j =$$

$$\left[P(j|x,\theta) - P(M|x,\theta)\frac{Q^*(j)}{Q^*(M)}\right] \Big/ \left[\sum_{j'=1}^{M} P(j'|x,\theta)\frac{H^*(j')}{Q^*(j')}\right]$$

First note that

$$(52)\quad \sum_{s=1}^{S} H_s \frac{\sum_{i' \in \mathcal{J}(t)} P(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')} = \sum_{j=1}^{M} P(j|x,\theta) H(j)/Q(j)$$

and a similar relation with $\frac{\partial P}{\partial \theta}(i|x,\theta)$ subsitituted for $P(i|x,\theta)$. After substituting (52) in (50), the latter is, when evaluated at $Q = Q^*$ and $H = H^*$, equal to (44). After substituting (52) in (51), $\tilde{\psi}_2$ is equal to $A\psi_2$, ψ_2 as in (43), with A equal to

$$A_{ii} = 1 - \frac{H(i)Q(M)^2}{H(M)Q(i)^2} \qquad A_{ij} = -\frac{H(j)Q(M)^2}{H(M)Q(i)^2} \quad \text{for } i \neq j$$

This shows that the moments used in Cosslett's estimator are a linear combination of those used in the new estimator. Hence, the covariance matrix of the latter cannot be larger than the covariance matrix of the former. The new estimator is easier to compute than Cosslett's estimator in this case. The optimization in the known Q and H case is only over the parameter β and that problem is much better behaved than the one where λ has to be estimated as well.

To compare the estimators for the unknown Q case consider first the moments (25)–(27). They are more difficult to compare to (42)–(44) than in

the previous case since they involve not only different parameters but also parameters of different dimension. λ is of dimension $S-1$, Q of dimension $M-1$. We will show that Cosslett's estimator cannot be better than the new estimator by adding moments, parameters and restrictions to (25)-(27) in such a way that the covariance matrix for $\hat{\theta}$ does not increase at each step, till we get the new estimator.

Consider the method of moments estimator for θ, H, λ and Q based on the moments (25)-(27) and (43). This does not change the covariance matrix of θ compared to the method of moments estimator for θ, H and λ based on (25)-(27). The only thing that has been changed is that $M-1$ parameters have been added together with $M-1$ moment equations. Now we add the $S-1$ restrictions $H_s/\lambda(s) = \sum_{i\in\mathcal{J}(s)} Q(i)$. This can only reduce the covariance matrix of θ. If we also make the substitution based on (52) we get the following moment equations, apart from (27) and (43) that do not change,

$$(53)\quad \tilde{\psi}_1(\theta,Q,H,s,i,x) = \frac{1}{P(i|x,\theta)}\frac{\partial P}{\partial\theta}(i|x,\theta)$$

$$-\Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)}\frac{\partial P}{\partial\theta}(i'|x,\theta)}{\sum_{i'\in\mathcal{J}(t)}Q(i')}\Big]\Big/\Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)}P(i'|x,\theta)}{\sum_{i'\in\mathcal{J}(t)}Q(i')}\Big]$$

$$(54)\quad \tilde{\psi}_2(\theta,Q,H,s,i,x)_s = \sum_{i'\in\mathcal{J}(s)} Q(i')$$

$$-\Big[\sum_{i'\in\mathcal{J}(s)} P(i'|x,\theta)\Big]\Big/\Big[\sum_{t=1}^{S} H_t \frac{\sum_{i'\in\mathcal{J}(t)}P(i'|x,\theta)}{\sum_{i'\in\mathcal{J}(t)}Q(i')}\Big]$$

(53) is equal to (44) and (54) is a linear combination of elements of (44). Hence the estimator based on moments (53), (54), (25) and (43), which is not worse than Cosslett's estimator, does not do better than the new estimator. That gives us the desired result that the PML estimator never does better than the new estimator.

4 Conclusion

In this paper an alternative estimation procedure is proposed for choice–based samples. In choice–based samples the sampling is conditional on the dependent variable. Therefore standard maximum likelihood techniques do not apply if only the conditional distribution of the dependent variable is specified. Various estimators have been proposed to deal with this problem. Some of them, the WESML and the CML estimators are not efficient. Cosslett's estimators are efficient but computationally demanding. All three of these estimators have the unusual feature that replacing some of the parameters by their probability limits reduces the efficiency with which the remaining parameters are estimated.

In the new estimation procedure some of the problems with the previously proposed estimators are solved. The new estimator is efficient while the computational burden is reduced compared with Cosslett's estimator. The case where the marginal probabilities of the choices are known and that where they are not known are both special cases of the general estimator. Efficiency is proven using recently developed concepts from Semiparametric estimation. This procedure also indicates a way of testing discrete choice models if the marginal choice–probabilities are known.

A Method of Moments Theory

First the apparatus for method of moments estimation will be set up. This account is based on Hansen [14] (without the complications of the dependence inherent in timeseries) and Manski [21].

Lemma A.1 *Let $h(z,\beta)$ be a function on $\mathcal{Z} \times \mathcal{B}$ with $\mathcal{Z}$ a Euclidean space and $\mathcal{B}$ a compact subset of a Euclidean space. Let h be a continuous function of β for all $z \in \mathcal{Z}$ and a measurable function of z for all $\beta \in \mathcal{B}$. Assume that $z_1, z_2, \ldots, z_N$ are independent, identically distributed random variables with distribution function F, and that $\|h(z,\beta)\| < g(z)$ for all $z \in \mathcal{Z}$ and $\beta \in \mathcal{B}$ for some g satisfying $\int g(z)dF(z) < \infty$. Then:*

$$\frac{1}{N}\sum_{n=1}^{N} h(z_n,\beta) \xrightarrow{\text{a.s.}} h(\beta) = \int h(z,\beta)dF(z) \quad \text{uniformly in } \mathcal{B}$$

and $h(\beta)$ is continous in β.

proof: see Jennrich [18]

Lemma A.2 *Let $h_N(\omega,\beta)$ be a function on $\Omega \times \mathcal{B}$ with Ω a measurable space and $\mathcal{B}$ a subset of a Euclidean space. Let $h_N(\omega,\beta)$ be a continous function of β for all $\omega \in \Omega$ and a measurable function of ω for all $\beta \in \mathcal{B}$. Then there exists a measurable function $\hat{\beta}_N(\omega)$ such that*

$$h_N(\omega,\hat{\beta}_N(\omega)) = \inf_{\beta} h_N(\omega,\beta)$$

for all $\omega \in \Omega$. If also $h_N(\omega,\beta) \longrightarrow h(\beta)$ almost surely, uniformly in β and if $h(\beta)$ has unique minimum at β^*, then:

$$\hat{\beta}_N(\omega) \xrightarrow{\text{a.s.}} \beta^*$$

proof: see Amemiya [1]

Assumption A.1 *$z_n(\omega)$, $n = 1, \ldots$ are random variables defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and take on values in $\mathcal{Z}$, a subspace of $\Re^L$. All z_n are identically distributed with induced distribution function F and independent.*

Assumption A.2 *$\mathcal{B}$ is the parameterspace, a compact subset of $\Re^K$. β^* is an element of the interior of $\mathcal{B}$*

Assumption A.3 *$h(z,\beta)$ is a continous function on $\mathcal{Z} \times \mathcal{B}$ to $\Re^M$ satisfying*

1. *$h(\beta) = \int h(z,\beta)dF(z) = 0$ implies $\beta = \beta^*$.*
2. *$\|h(z,\beta)\| < g(z)\ \forall z \in \mathcal{Z}, \beta \in \mathcal{B}$ with $\int g(z)dF(z) < \infty$*

Assumption A.4 *$C_N(\omega) \xrightarrow{a.s.} C_0$ as $N \longrightarrow \infty$ with C_0 a positive definite, symmetric $M \times M$ matrix.*

Theorem A.3 (consistency) *Suppose assumptions 1-4 hold. Then, the estimator $\hat{\beta}_N(\omega)$ for β^*, defined by*

$$\mathrm{R}_N(\omega, \hat{\beta}_N(\omega)) = \inf_{\beta} \mathrm{R}_N(\omega, \beta)$$

with

$$\mathrm{R}_N(\omega,\beta) = \left[\frac{1}{N}\sum_{n=1}^{N} h(z_n(\omega),\beta)\right]' \cdot C_N(\omega) \cdot \left[\frac{1}{N}\sum_{n=1}^{N} h(z_n(\omega),\beta)\right]$$

satisfies

$$\hat{\beta}_N(\omega) \xrightarrow{a.s.} \beta^*$$

proof: The first step is to note that we cann apply lemma 1 to prove uniform convergence of

$$\frac{1}{N}\sum_{n=1}^{N} h(z_n(\omega),\beta) \xrightarrow{a.s.} h(\beta) = \int h(z,\beta)dF(z)$$

This implies uniform convergence of

$$\mathrm{R}_N(\omega,\beta) \xrightarrow{a.s.} h(\beta)' \cdot C_0 \cdot h(\beta)$$

Because $h(\beta)$ has a unique zero at $\beta = \beta^*$, and because of the positive definiteness of C_0, the limit of R_N has a unique minimum at the same value. This ensures that we can apply lemma 2 to get the required result. $\mathcal{QED}$.

Assumption A.5 *Let $h(z,\beta)$ be continously differentiable in β and let its derivative with respect to β be a measurable function of z for all $\beta \in \mathcal{B}$.*

Define the matrix $\Gamma(\beta)$ as:

$$\Gamma(\beta) = \int \frac{\partial h}{\partial \beta'}(z,\beta) dF(z)$$

Assumption A.6 *For any sequence β_N, converging almost surely to β^*, the matrix $\Gamma(\beta_N)$ converges almost surely to*

$$\Gamma_0 = \Gamma(\beta^*)$$

and Γ_0 has full rank.

Theorem A.4 (Normality) *Suppose assumptions 1-6 hold. Then $\hat{\beta}_N(\omega)$ satisfies:*

$$\sqrt{N}(\hat{\beta}_N(\omega) - \beta^*) \xrightarrow{d} \mathcal{N}\left(0, (\Gamma_0' C_0 \Gamma_0)^{-1} \Gamma_0' C_0 \Delta_0 C_0 \Gamma_0 (\Gamma_0' C_0 \Gamma_0)^{-1}\right)$$

where $\Delta_0 = \int h(z,\beta^*) \cdot h(z,\beta^*)' dF(z)$

proof: Theorem 1 guarantees that $\hat{\beta}_N(\omega) \xrightarrow{a.s.} \beta^*$. Since $\beta^* \in \text{int}(\mathcal{B})$, for large enough N the estimator $\hat{\beta}_N(\omega)$ must be in the interior of $\mathcal{B}$, with probability one. Then it must satisfy the first order conditions for a minimum:

$$0 = \frac{\partial h_N}{\partial \beta}(\omega,\beta) \cdot C_N \cdot h_N(\omega,\beta)$$

where we use the shorthand $h_N(\omega,\beta) = \frac{1}{N}\sum_{n=1}^{N} h(z_n(\omega),\beta)$.

Apply a mean value theorem to the j$^{\text{th}}$ element of $h_N(\omega,\beta)$:

$$h_N(\omega,\hat{\beta}_N)_j = h_N(\omega,\beta^*)_j + \frac{\partial h_N}{\partial \beta}(\omega,\tilde{\beta}^j)_j \cdot (\hat{\beta}_N - \beta^*)$$

where $\tilde{\beta}^j = \beta^* + \lambda^j \cdot (\hat{\beta}_N - \beta^*)$ for some $0 \leq \lambda^j \leq 1$.
Define: $\tilde{\Gamma}_N$ to be the matrix with j$^{\text{th}}$ row equal to:

$$\tilde{\Gamma}_{N,j} = \frac{\partial h_N}{\partial \beta'}(\omega,\tilde{\beta}^j)_j$$

This enables us to write the first order conditions as

$$\Gamma'(\hat{\beta}_N) \cdot C_N \cdot \left[h_N(\omega, \beta^*) + \tilde{\Gamma}_N(\hat{\beta}_N - \beta^*) \right] = 0$$

The continuity conditions on h and the almost sure convergence of $\hat{\beta}_N \xrightarrow{a.s.} \beta^*$ and therefore of $\tilde{\beta}^j \xrightarrow{a.s.} \beta^*$ guarantee:

$$\Gamma(\hat{\beta}_N) \xrightarrow{a.s.} \Gamma_0 \quad \tilde{\Gamma}_N \xrightarrow{a.s.} \Gamma_0$$

Because of the full rank of Γ_0 and the invertibility of C_0, we can, for large enough N, with probability one, write the first order conditions as:

$$\sqrt{N}(\hat{\beta}_N - \beta^*) = \left(\Gamma'(\hat{\beta}_N) C_N \tilde{\Gamma}_N \right)^{-1} \Gamma'(\hat{\beta}_N) C_N \sqrt{N} h_N(\omega, \beta^*)$$

Because of the independence of the z_n and the identical distribution, we can use a Central Limit Theorem to prove that

$$\sqrt{N} h_N(\omega, \beta^*) \xrightarrow{d} \mathcal{N}(0, \Delta_0)$$

The remainder follows from the almost sure convergence of all the other factors in the first order conditions. $\mathcal{QED}$.

If Δ_0 is nonsingular the optimal choice for C_0, the limit of the weight matrix, is Δ_0^{-1}. In that case the covariance matrix reduces to $(\Gamma_0' \Delta_0^{-1} \Gamma_0)^{-1}$. If Δ_0 is singular it can be written by rearranging the moment equations as the following partioned matrix:

$$\Delta_0 = \begin{pmatrix} \Delta_{011} & \Delta_{012} \\ \Delta_{021} & \Delta_{022} \end{pmatrix}$$

with $dim(\Delta_{011}) = rank(\Delta_{011}) = rank(\Delta_0)$. The optimal C_0 is then

$$C_0 = \begin{pmatrix} \Delta_{011}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

Even though C_0 is singular in that case, as long as $\Gamma_0' C_0 \Gamma_0$ is not, there is no problem.

B Theorems 2.1, 3.1, 3.2

proof of theorem 2.1:

All the estimators defined by

$$\sum_{n=1}^{N} w(i_n, s_n) \frac{1}{P(i_n|x_n, \hat{\theta})} \frac{\partial P}{\partial \theta}(i_n|x_n, \hat{\theta}) = 0$$

with $w(i,s)$ satisfying

$$\sum_{s|i \in \mathcal{J}(s)} \frac{H_s^*}{\sum_{i' \in \mathcal{J}(s)} Q(i')} w(i,s) = 1$$

are method of moment estimators. Their consistency and asymptotic normality follows from the theorems in appendix A. The asymptotic covariance matrix of the typical element of this matrix is

$$\Sigma_w = \Gamma_w^{-1} \Delta_w \Gamma_w'^{-1}$$

with

$$\Delta_w = \mathrm{E}_g w(i,s)^2 \Big[\frac{1}{P(i|x,\theta^*)^2} \frac{\partial P}{\partial \theta}(i|x,\theta^*) \frac{\partial P}{\partial \theta'}(i|x,\theta^*) \Big]$$

and

$$\Gamma_w = \mathrm{E}_g w(i,s) \cdot$$

$$\Big[\frac{1}{P(i|x,\theta^*)} \frac{\partial^2 P}{\partial \theta \partial \theta'}(i|x,\theta^*) - \frac{1}{P(i|x,\theta^*)^2} \frac{\partial P}{\partial \theta}(i|x,ts) \frac{\partial P}{\partial \theta'}(i|x,\theta^*) \Big]$$

The expectation of $w(i,s0$ given i is $Q(i)/H(i)$ for all weighting schemes in this class. Hence, Γ_w is the same for all weighting schemes. The scheme that minimizes Δ is that where $w(i,s)$ is equal to its expectation given i. Thus optimally $w(i,s) = Q(i)/H(i)$. $\mathcal{QED}$

proof of lemma 1:

Suppose the likelihood function with N observations z_n is $L(\beta) = \sum_{n=1}^{N} \ln f(z_n, \beta)$. The asymptotic covariance matrix V of $\sqrt{N}(\hat{\beta} - \beta^*)$ is:

$$V = I(\beta) = \Big[\mathrm{E} \frac{\partial \ln f}{\partial \beta}(z, \beta^*) \cdot \frac{\partial \ln f}{\partial \beta'}(x, \beta^*) \Big]^{-1}$$

Partition V and its inverse V^{-1} according to β_1, β_2 and β_3:

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix} \quad V^{-1} = \begin{pmatrix} V^{11} & V^{12} & V^{13} \\ V^{21} & V^{22} & V^{23} \\ V^{31} & V^{32} & V^{33} \end{pmatrix}$$

The variance of the constrained estimator of β_1 and β_3 given $\beta_2 = 0$ is:

$$\begin{pmatrix} V^{11} & V^{13} \\ V^{31} & V^{33} \end{pmatrix}^{-1} = \begin{pmatrix} \left(V^{11} - V^{13}(V^{33})^{-1}V^{31}\right)^{-1} & \cdots \\ \cdots & \cdots \end{pmatrix}$$

Since we could characterize the maximum likelihood estimates of β_1 and β_2 by

$$\sum_{n=1}^{N} h_1(\hat{\beta}_1, \hat{\beta}_2, z_n) = 0$$

the covariance matrix must satisfy:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \left[\left[\mathrm{E}\ \frac{\partial h_1}{\partial(\beta_1' \ \beta_2')}\right]'\left[\mathrm{E}\ h_1 h_1'\right]^{-1}\left[\mathrm{E}\ \frac{\partial h_1}{\partial(\beta_1' \ \beta_2')}\right]\right]^{-1}$$

The estimator for β_1 given $\beta_2 = 0$ based on minimization of

$$\frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1, 0, z_n)' \cdot C_N \cdot \frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1, 0, z_n)$$

with $C_N \xrightarrow{a.s.} \mathrm{E}h_1 h_1'$ has asymptotic covariance matrix:

$$\left[\left[\mathrm{E}\ \frac{\partial h_1}{\partial \beta_1'}\right]'\left[\mathrm{E}\ h_1 h_1'\right]^{-1}\left[\mathrm{E}\ \frac{\partial h_1}{\partial \beta_{[}1}'\right]\right]^{-1}$$

$$= \left[\left(V_{11} - V_{12}V_{22}^{-1}V_{21}\right)^{-1}\right]^{-1} = V_{11} - V_{12}V_{22}^{-1}V_{21}$$

This is equal to $[V^{11} - V^{13}(V^{33})^{-1}V^{31}]^{-1}$ after some rearranging. $\mathcal{QED}$.

proof of theorem 3.2:

The assumptions made, (2.1)–(2.4) and (3.3) guarantee the assumptions needed for thorem A.3 and A.4 in the first appendix to hold. $\mathcal{QED}$.

C Local Asymptotic Minimax Efficiency

In the section 3.1 we found an estimator that achieves the Cramér–Rao lower bound. This would be sufficient if one is satisfied with the assumption of discreteness of x. One might argue that all one ever has are discrete data. This appendix is devoted to the extension to continous data to make precise the way in which the estimator does or does not depend on discreteness of the regressors.

To analyse efficiency we have to either define classes of estimators that either exclude superefficient estimators[8] or consider criteria that penalize them. Otherwise standard maximum likelihood estimators would not be efficient and we cannot hope to have efficiency for the estimator proposed in the section 3.1. The two approaches have in common that they look at the behaviour of the estimators if the parameter varies in the neighbourhood of the true value. An example of the first approach is the concept of regular estimators. Regular estimators are estimators $\hat{\beta}_N$ for which the asymptotic distribution of $\sqrt{N}(\hat{\beta}_N - \beta_N)$ does not depend on the particular sequence β_N, provided the latter converges to β^*. See for a discussion Newey [24] and Begun, Hall, Huang and Wellner [4].

The approach followed here is the second one. We look at the expected loss of a particular estimator as the true value of the parameter β varies over the neighbourhood $N_\delta(\beta^*)$ defined as $N_\delta(\beta^*) = \{\beta | \|\beta - \beta^*\| < \delta\}$. In particular the worst case (i.e. the worst possible value of β in this neighbourhood) is considered. It is the expected loss in this worst case that we try to minimize over the space of all estimators. Superefficient estimators usually do quite badly in the neighbourhood of the true value of the parameter and therefore they will have a high expected maximum loss. This approach was proposed by Hájek [13]. It was applied in a method of moments context by Chamberlain [5,6].

The first step is to define the class of loss functions considered. If the method were very sensitive to the particular loss function used it would of course be of less interest. Fortunately this turns out not to be the case. A loss function $\ell : \Re \to \Re$ is an element of the class of loss functions $\mathcal{L}$ if:

1. for all $u \in \Re$ $\ell(u) = \ell(\|u\|)$

[8] see Newey [24] page 4 for an example of superefficient estimators

2. for all $u, v \in \Re$ $\|u\| \geq \|v\|$ implies $\ell(u) \geq \ell(v)$

3. $\int_{-\infty}^{\infty} \ell(u) \exp(-\lambda u^2/2) du < \infty$ for $\lambda > 0$

4. $\ell(0) = 0$

Consider the family of probability functions $f(z, \beta)$ of a discrete random variable z with finite support $(z^1, z^2, \ldots, z^L)$ for $\beta \in \mathcal{B}$, a subset of $\Re^K$. Consider a sequence h_N, converging to $h \in \Re^K$, and the sequence $\beta_N = \beta + N^{-1/2} \cdot h_N$ such that $\beta_N \in \mathcal{B}$ for all N. We are interested in the asymptotic behaviour of the likelihood ratio

$$(55) \quad L_N(\beta, h) = \sum_{n=1}^{N} \ln \frac{f(z_n, \beta_N)}{f(z_n, \beta)}$$

where the z_n are independent draws from $f(z, \beta)$

LeCam [20] shows that under standard assumptions $L_N(\beta, h)$ has asymptotically a normal distribution with parameters depending on h and the information matrix. The following version of the *local asymptotic normality* (LAN) condition does not give the weakest set of conditions, but they are easy to check and follow from other assumptions already made.

Lemma C.1 *assume that $f(z, \beta)$ is twice continously differentiable with respect to β for all $\beta \in \mathcal{B}$ and satisfies the Information Matrix equality. Then, for all $\beta \in \mathcal{B}$, the likelihood ratio $L_N(\beta, h)$ satisfies:*

$$L_N(\beta, h) \xrightarrow{d} \mathcal{N}\left(-\frac{1}{2} h' I(\beta) h, h' I(\beta) h\right)$$

where $I(\beta)$ is the information matrix:

$$I(\beta) = \int f(z, \beta) \frac{\partial \ln f}{\partial \beta}(z, \beta) \cdot \frac{\partial \ln f}{\partial \beta'}(z, \beta) dz$$

Proof: see LeCam [20]

As leCam argues, the conditions sufficient for LAN are usually implied by sufficient conditions for asymptotic normality for maximum likelihood estimators.

The next step is to apply a theorem by Hájek [13] that shows that under LAN maximum likelihood estimators have particular desirable properties:

Lemma C.2 *If local asymptotic normality holds, then the maximum risk associated with an estimator T_N of the first[9] element β_1^* of a parameter vector β^* is bounded from below in the following way:*

$$\lim_{\delta \downarrow 0} \lim_{N \to \infty} \inf_{T_N} \sup_{\beta \in N_\delta(\beta^*)} \mathrm{E}_{f(z,\beta^*)} \ell(\sqrt{N}(T_N - \beta_1)) \geq$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \ell(\sigma u) \exp(-u^2/2) du$$

where σ is the square root of the (1,1) element of the inverse of the information matrix at β^*, $I(\beta^*)$

Proof: see Hajek [13]

If the maximum likelihood estimator is asymptotically normal its maximum expected loss would be equal to

$$\mathrm{E}_{f(z,\beta^*)} \ell(\sqrt{N}(T_N - \beta_1)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \ell(u) \exp(-u^2/2\sigma^2) du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \ell(\sigma u) \exp(-u^2/2) du$$

So one knows that the lower bound on the maximum expected loss can be attained. One can therefore interpret the lemma as stating that standard maximum likelihood estimators cannot be beaten asymptotically if one varies the parameter value over an increasingly small neighbourhood of the true value. If an estimator T_N has a lower risk at a particular value of the parameter space β, then it must do worse than the maximum likelihood estimator for another value of the parameter, arbitrarily close to β.

Now it is time to apply these concepts to the estimators analyzed in the section 3.1. The only conditions that have to be checked in these discrete models with finite support is that the probabilities are twice continuously differentiable with respect to the underlying parameters. Here it turns out to be convenient that by assuming discreteness for the x's, the whole model becomes discrete. First the model with parameters (H θ π) and associated likelihood (31) will be checked. The probability of an observation (s, i, l) is

[9] concentrating on the first element of β does not involve any loss of generality since the parameters can be reordered

$$(56)\quad H_s \frac{P(i|x_l,\theta)\pi_l}{\sum_{i'\in\mathcal{J}(s)}\sum_{m=1}^{L}\pi_m P(i'|x_m,\theta)}$$

Because of assumption (2.2) this is a twice continously differentiable function of θ, H and π. It therefore satisfies the LAN condition and the lower bound on the maximum risk applies.

For the transformed model with the parameters (H Q θ π_2) the same result holds. Define $\tilde{\pi} = (\tilde{\pi}_1' \ \tilde{\pi}_2')'$ with $\tilde{\pi}_2 = \pi_2$ and

$$(57)\quad \tilde{\pi}_1 = V_0^{-1}(\theta)(Q - V_1(\theta)\pi_2)$$

$\tilde{\pi}_1$ is a twice continously differentiable function of Q, θ and π_2. Therefore the probability of an observation (s,i,l), now equal to:

$$(58)\quad H_s \frac{P(i|x_l,\theta)\tilde{\pi}_l}{\sum_{i'\in\mathcal{J}(s)} Q(i')}$$

satisfies the conditions for lemma C.1. The LAN condition holds again and therefore the bound applies.

The case with linear restrictions on H, Q and θ is trivial. Continuity of the first two derivatives follows from the above analysis, and the bound applies.

What has been shown now is that the estimators define in section 3.1 have particular properties. These properties hold for all well behaved maximum likelihood estimators so it is not so surprising that they do so for the estimators under consideration. The importance of this is that it enables one to compare estimators some of which do not fit in the maximum likelihood framework. These estimators will still have a well defined risk. One can compare this risk to that for models that despite being very close to the ones studied, do satisfy the conditions for maximum likelihood theory to apply.

We want to apply the efficiency result to the case where x does not have finite support. The procedure we follow is based on Chamberlain [5]. The main result needed is:

Lemma C.3 *Let $h : \mathcal{Z} \to \Re^t$ be a measurable function and let F be a probability measure with support $\mathcal{Z}_F \subset \mathcal{Z}$. If $\int ||h|| dF < \infty$ then there exists a probability measure G whose support is a finite subset of $\mathcal{Z}_F$ and which satisfies $\int h dG = \int h dF$*

proof: see Chamberlain [5]

proof of theorem 3.3

We apply lemma C.3 to the measure with associated distribution function R of x with as h function the moments $\tilde{\psi}$, $\frac{\partial \tilde{\psi}}{\partial \gamma}$, $\tilde{\psi} \cdot \tilde{\psi}'$ and the b_j from the particular neighbourhood definition. The lemma then implies that in any neighbourhood of F (i.e. for any given, finite set of b_j) we can find a discrete measure that satisfies the same moment restrictions (the $\tilde{\psi}$), and also has the same expected derivatives and outer products of these moments (the $\frac{\partial \tilde{\psi}}{\partial \gamma}$ and the $\tilde{\psi} \cdot \tilde{\psi}'$). Since we know that the expected loss for such model with discrete support was bounded by lemma 3.3, this bound also applies to the case of non-discrete measures. $\mathcal{QED}$.

References

[1] Amemiya, T., "Regression Analysis when the Dependent Variable is Truncated Normal", *Econometrica*, vol 41, 997–1017, 1973

[2] Amemiya, T., *Advanced Econometrics*, Harvard University Press, Cambridge, MA, 1987

[3] Amemiya, T., and Q. H. Vuong, "A Comparison of Two Consistent Estimators in the Choice-based Sampling Qualitative Response Model", *Econometrica*, vol 55, 699–702, 1987

[4] Begun, J. M., W. J. Hall, W–M. Huang and J. A. Wellner, "Information and Asymptotic Efficiency in Parametric–Nonparametric models", *Annals of Statistics*, vol 11, 432–452, 1983

[5] Chamberlain, G., "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, vol 34, 305–334, 1987

[6] Chamberlain, G., "Efficiency Bounds for Semi–Parametric Regression", working paper, department of economics, University of Wisconsin, 1987

[7] Cosslett, S. R., "An Efficient Estimator of Discrete–Choice Models for Choice–Based Samples", working paper, department of economics, University of California, Berkeley, 1978

[8] Cosslett, S. R., "Maximum Likelihood Estimation for Choice–based Samples", *Econometrica*, vol 49, 1289–1316, 1981

[9] Cosslett, S. R., "Efficient Estimation of Discrete Choice Models", in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, 51–111, MIT Press, Cambridge, MA, 1981

[10] Engle, R., D. Hendry, and J. F. Richard, "Exogeneity", *Econometrica*, vol 51, 277–304, 1983

[11] Gill, R. D., Y. Vardi, and J. A. Wellner, "Large Sample Theory of Empirical Distributions in Biased Sampling Models", *The Annals of Statistics*, vol 16, 1069–1112, 1988

[12] Gourieroux, C., and A. Montfort, "Econometrics Based on Endogenous Samples", working paper, 1989

[13] Hájek, J., "Local Asymptotic Minimax and Admissability in Estimation", *proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, University of California Press, Berkeley, CA 1972

[14] Hansen, L. P., "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, vol 50, 1029–1054, 1982

[15] Hoem, J., "weet ik niet", in J. J. Heckman and B. Singer eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge, 1984

[16] Heckman, J. J., and V. J. Hotz, "Choosing Among Alternative Nonexperimental methods for Estimating the Impact of Social Programs: The Case of Manpower Training", discussion paper, Economics Research Center, NORC, 1988

[17] Hsieh, D. A., C. F. Manski and D. McFadden, "Estimation of Response probabilities from Augmented Retrospective Observations", *Journal of the American Statistical Association* vol 80, 651–662, 1985

[18] Jennrich, R. I., "Asymptotic Properties of Nonlinear Least Squares Estimators", *The Annals of Mathematical Statistics*, vol 40, 633–644, 1969

[19] Lancaster, T., and G. Imbens, "Choice-Based Sampling of Dynamic Populations", forthcoming in: Hartog, Ridder and Theeuwes, eds., *Panel Data and Labor Market Studies*, 1989

[20] LeCam, L., "On the Assumptions used to prove asymptotic normality of maximum likelihood estimates", *Annals of Mathematical Statistics*, vol 41, 802–828, 1970

[21] Manski, C. F., *Analog Estimation Methods in Econometrics*, Chapman and Hall, New York, NY, 1988

[22] Manski, C. F., and S. R. Lerman, "The Estimation of Choice Probabilities from Choice-based Samples", *Econometrica*, vol 45, 1977-1988, 1977

[23] Manski, C. F., and D. McFadden, "Alternative Estimators and Sample Designs for Discrete Choice Analysis", in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, 51–111, MIT Press, Cambridge, MA, 1981

[24] Newey, W., "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, vol 14, 201–206, 1984

[25] Newey, W., "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, vol 53, 1047–1069, 1985

[26] Newey, W., "An introduction to Semiparametric Efficiency Bounds", working paper, Princeton University, 1988

[27] Ridder, G., *Life Cycle Patterns in Labor Market Experience*, Phd dissertation, University of Amsterdam, 1987

[28] Xie, Y., and C. F. Manski, "The logit model and response–based sampling", working paper 8811, social systems research institute, University of Wisconsin, 1988

IN 1988 REEDS VERSCHENEN

297 Bert Bettonvil
Factor screening by sequential bifurcation

298 Robert P. Gilles
On perfect competition in an economy with a coalitional structure

299 Willem Selen, Ruud M. Heuts
Capacitated Lot-Size Production Planning in Process Industry

300 J. Kriens, J.Th. van Lieshout
Notes on the Markowitz portfolio selection method

301 Bert Bettonvil, Jack P.C. Kleijnen
Measurement scales and resolution IV designs: a note

302 Theo Nijman, Marno Verbeek
Estimation of time dependent parameters in lineair models
using cross sections, panels or both

303 Raymond H.J.M. Gradus
A differential game between government and firms: a non-cooperative approach

304 Leo W.G. Strijbosch, Ronald J.M.M. Does
Comparison of bias-reducing methods for estimating the parameter in dilution series

305 Drs. W.J. Reijnders, Drs. W.F. Verstappen
Strategische bespiegelingen betreffende het Nederlandse kwaliteits-concept

306 J.P.C. Kleijnen, J. Kriens, H. Timmermans and H. Van den Wildenberg
Regression sampling in statistical auditing

307 Isolde Woittiez, Arie Kapteyn
A Model of Job Choice, Labour Supply and Wages

308 Jack P.C. Kleijnen
Simulation and optimization in production planning: A case study

309 Robert P. Gilles and Pieter H.M. Ruys
Relational constraints in coalition formation

310 Drs. H. Leo Theuns
Determinanten van de vraag naar vakantiereizen: een verkenning van materiële en immateriële factoren

311 Peter M. Kort
Dynamic Firm Behaviour within an Uncertain Environment

312 J.P.C. Blanc
A numerical approach to cyclic-service queueing models

313 Drs. N.J. de Beer, Drs. A.M. van Nunen, Drs. M.O. Nijkamp
Does Morkmon Matter?

314 Th. van de Klundert
Wage differentials and employment in a two-sector model with a dual labour market

315 Aart de Zeeuw, Fons Groot, Cees Withagen
On Credible Optimal Tax Rate Policies

316 Christian B. Mulder
Wage moderating effects of corporatism
Decentralized versus centralized wage setting in a union, firm, government context

317 Jörg Glombowski, Michael Krüger
A short-period Goodwin growth cycle

318 Theo Nijman, Marno Verbeek, Arthur van Soest
The optimal design of rotating panels in a simple analysis of variance model

319 Drs. S.V. Hannema, Drs. P.A.M. Versteijne
De toepassing en toekomst van public private partnership's bij de grote en middelgrote Nederlandse gemeenten

320 Th. van de Klundert
Wage Rigidity, Capital Accumulation and Unemployment in a Small Open Economy

321 M.H.C. Paardekooper
An upper and a lower bound for the distance of a manifold to a nearby point

322 Th. ten Raa, F. van der Ploeg
A statistical approach to the problem of negatives in input-output analysis

323 P. Kooreman
Household Labor Force Participation as a Cooperative Game; an Empirical Model

324 A.B.T.M. van Schaik
Persistent Unemployment and Long Run Growth

325 Dr. F.W.M. Boekema, Drs. L.A.G. Oerlemans
De lokale produktiestructuur doorgelicht.
Bedrijfstakverkenningen ten behoeve van regionaal-economisch onderzoek

326 J.P.C. Kleijnen, J. Kriens, M.C.H.M. Lafleur, J.H.F. Pardoel
Sampling for quality inspection and correction: AOQL performance criteria

327 Theo E. Nijman, Mark F.J. Steel
Exclusion restrictions in instrumental variables equations

328 B.B. van der Genugten
Estimation in linear regression under the presence of heteroskedasticity of a completely unknown form

329 Raymond H.J.M. Gradus
The employment policy of government: to create jobs or to let them create?

330 Hans Kremers, Dolf Talman
Solving the nonlinear complementarity problem with lower and upper bounds

331 Antoon van den Elzen
Interpretation and generalization of the Lemke-Howson algorithm

332 Jack P.C. Kleijnen
Analyzing simulation experiments with common random numbers, part II: Rao's approach

333 Jacek Osiewalski
Posterior and Predictive Densities for Nonlinear Regression.
A Partly Linear Model Case

334 A.H. van den Elzen, A.J.J. Talman
A procedure for finding Nash equilibria in bi-matrix games

335 Arthur van Soest
Minimum wage rates and unemployment in The Netherlands

336 Arthur van Soest, Peter Kooreman, Arie Kapteyn
Coherent specification of demand systems with corner solutions and endogenous regimes

337 Dr. F.W.M. Boekema, Drs. L.A.G. Oerlemans
De lokale produktiestruktuur doorgelicht II. Bedrijfstakverkenningen ten behoeve van regionaal-economisch onderzoek. De zeescheepsnieuwbouwindustrie

338 Gerard J. van den Berg
Search behaviour, transitions to nonparticipation and the duration of unemployment

339 W.J.H. Groenendaal and J.W.A. Vingerhoets
The new cocoa-agreement analysed

340 Drs. F.G. van den Heuvel, Drs. M.P.H. de Vor
Kwantificering van ombuigen en bezuinigen op collectieve uitgaven 1977-1990

341 Pieter J.F.G. Meulendijks
An exercise in welfare economics (III)

342 W.J. Selen and R.M. Heuts
A modified priority index for Günther's lot-sizing heuristic under capacitated single stage production

343 Linda J. Mittermaier, Willem J. Selen, Jeri B. Waggoner, Wallace R. Wood
Accounting estimates as cost inputs to logistics models

344 Remy L. de Jong, Rashid I. Al Layla, Willem J. Selen
Alternative water management scenarios for Saudi Arabia

345 W.J. Selen and R.M. Heuts
Capacitated Single Stage Production Planning with Storage Constraints and Sequence-Dependent Setup Times

346 Peter Kort
The Flexible Accelerator Mechanism in a Financial Adjustment Cost Model

347 W.J. Reijnders en W.F. Verstappen
De toenemende importantie van het verticale marketing systeem

348 P.C. van Batenburg en J. Kriens
E.O.Q.L. - A revised and improved version of A.O.Q.L.

349 Drs. W.P.C. van den Nieuwenhof
Multinationalisatie en coördinatie
De internationale strategie van Nederlandse ondernemingen nader beschouwd

350 K.A. Bubshait, W.J. Selen
Estimation of the relationship between project attributes and the implementation of engineering management tools

351 M.P. Tummers, I. Woittiez
A simultaneous wage and labour supply model with hours restrictions

352 Marco Versteijne
Measuring the effectiveness of advertising in a positioning context with multi dimensional scaling techniques

353 Dr. F. Boekema, Drs. L. Oerlemans
Innovatie en stedelijke economische ontwikkeling

354 J.M. Schumacher
Discrete events: perspectives from system theory

355 F.C. Bussemaker, W.H. Haemers, R. Mathon and H.A. Wilbrink
A (49,16,3,6) strongly regular graph does not exist

356 Drs. J.C. Caanen
Tien jaar inflatieneutrale belastingheffing door middel van vermogensaftrek en voorraadaftrek: een kwantitatieve benadering

357 R.M. Heuts, M. Bronckers
A modified coordinated reorder procedure under aggregate investment and service constraints using optimal policy surfaces

358 B.B. van der Genugten
Linear time-invariant filters of infinite order for non-stationary processes

359 J.C. Engwerda
LQ-problem: the discrete-time time-varying case

360 Shan-Hwei Nienhuys-Cheng
Constraints in binary semantical networks

361 A.B.T.M. van Schaik
Interregional Propagation of Inflationary Shocks

362 F.C. Drost
How to define UMVU

363 Rommert J. Casimir
Infogame users manual
Rev 1.2 December 1988

364 M.H.C. Paardekooper
A quadratically convergent parallel Jacobi-process for diagonal dominant matrices with nondistinct eigenvalues

365 Robert P. Gilles, Pieter H.M. Ruys
Characterization of Economic Agents in Arbitrary Communication Structures

366 Harry H. Tigelaar
Informative sampling in a multivariate linear system disturbed by moving average noise

367 Jörg Glombowski
Cyclical interactions of politics and economics in an abstract capitalist economy

IN 1989 REEDS VERSCHENEN

368 Ed Nijssen, Will Reijnders
"Macht als strategisch en tactisch marketinginstrument binnen de distributieketen"

369 Raymond Gradus
Optimal dynamic taxation with respect to firms

370 Theo Nijman
The optimal choice of controls and pre-experimental observations

371 Robert P. Gilles, Pieter H.M. Ruys
Relational constraints in coalition formation

372 F.A. van der Duyn Schouten, S.G. Vanneste
Analysis and computation of (n,N)-strategies for maintenance of a two-component system

373 Drs. R. Hamers, Drs. P. Verstappen
Het company ranking model: a means for evaluating the competition

374 Rommert J. Casimir
Infogame Final Report

375 Christian B. Mulder
Efficient and inefficient institutional arrangements between governments and trade unions; an explanation of high unemployment, corporatism and union bashing

376 Marno Verbeek
On the estimation of a fixed effects model with selective non-response

377 J. Engwerda
Admissible target paths in economic models

378 Jack P.C. Kleijnen and Nabil Adams
Pseudorandom number generation on supercomputers

379 J.P.C. Blanc
The power-series algorithm applied to the shortest-queue model

380 Prof. Dr. Robert Bannink
Management's information needs and the definition of costs, with special regard to the cost of interest

381 Bert Bettonvil
Sequential bifurcation: the design of a factor screening method

382 Bert Bettonvil
Sequential bifurcation for observations with random errors

383 Harold Houba and Hans Kremers
Correction of the material balance equation in dynamic input-output models

384 T.M. Doup, A.H. van den Elzen, A.J.J. Talman
Homotopy interpretation of price adjustment processes

385 Drs. R.T. Frambach, Prof. Dr. W.H.J. de Freytas
Technologische ontwikkeling en marketing. Een oriënterende beschouwing

386 A.L.P.M. Hendrikx, R.M.J. Heuts, L.G. Hoving
Comparison of automatic monitoring systems in automatic forecasting

387 Drs. J.G.L.M. Willems
Enkele opmerkingen over het inversificerend gedrag van multinationale ondernemingen

388 Jack P.C. Kleijnen and Ben Annink
Pseudorandom number generators revisited

389 Dr. G.W.J. Hendrikse
Speltheorie en strategisch management

390 Dr. A.W.A. Boot en Dr. M.F.C.M. Wijn
Liquiditeit, insolventie en vermogensstructuur

391 Antoon van den Elzen, Gerard van der Laan
Price adjustment in a two-country model

392 Martin F.C.M. Wijn, Emanuel J. Bijnen
Prediction of failure in industry
An analysis of income statements

393 Dr. S.C.W. Eijffinger and Drs. A.P.D. Gruijters
On the short term objectives of daily intervention by the Deutsche Bundesbank and the Federal Reserve System in the U.S. Dollar - Deutsche Mark exchange market

394 Dr. S.C.W. Eijffinger and Drs. A.P.D. Gruijters
On the effectiveness of daily interventions by the Deutsche Bundesbank and the Federal Reserve System in the U.S. Dollar - Deutsche Mark exchange market

395 A.E.M. Meijer and J.W.A. Vingerhoets
Structural adjustment and diversification in mineral exporting developing countries

396 R. Gradus
About Tobin's marginal and average q
A Note

397 Jacob C. Engwerda
On the existence of a positive definite solution of the matrix equation $X + A^T X^{-1} A = I$

398 Paul C. van Batenburg and J. Kriens
Bayesian discovery sampling: a simple model of Bayesian inference in auditing

399 Hans Kremers and Dolf Talman
Solving the nonlinear complementarity problem

400 Raymond Gradus
Optimal dynamic taxation, savings and investment

401 W.H. Haemers
Regular two-graphs and extensions of partial geometries

402 Jack P.C. Kleijnen, Ben Annink
Supercomputers, Monte Carlo simulation and regression analysis

403 Ruud T. Frambach, Ed J. Nijssen, William H.J. Freytas
Technologie, Strategisch management en marketing

404 Theo Nijman
A natural approach to optimal forecasting in case of preliminary observations

405 Harry Barkema
An empirical test of Holmström's principal-agent model that tax and signally hypotheses explicitly into account

406 Drs. W.J. van Braband
De begrotingsvoorbereiding bij het Rijk

407 Marco Wilke
Societal bargaining and stability

408 Willem van Groenendaal and Aart de Zeeuw
Control, coordination and conflict on international commodity markets

409 Prof. Dr. W. de Freytas, Drs. L. Arts
Tourism to Curacao: a new deal based on visitors' experiences

410 Drs. C.H. Veld
The use of the implied standard deviation as a predictor of future stock price variability: a review of empirical tests

411 Drs. J.C. Caanen en Dr. E.N. Kertzman
Inflatieneutrale belastingheffing van ondernemingen

412 Prof. Dr. B.B. van der Genugten
A weak law of large numbers for m-dependent random variables with unbounded m

413 R.M.J. Heuts, H.P. Seidel, W.J. Selen
A comparison of two lot sizing-sequencing heuristics for the process industry

414 C.B. Mulder en A.B.T.M. van Schaik
Een nieuwe kijk op structuurwerkloosheid

415 Drs. Ch. Caanen
De hefboomwerking en de vermogens- en voorraadaftrek

416 Guido W. Imbens
Duration models with time-varying coefficients